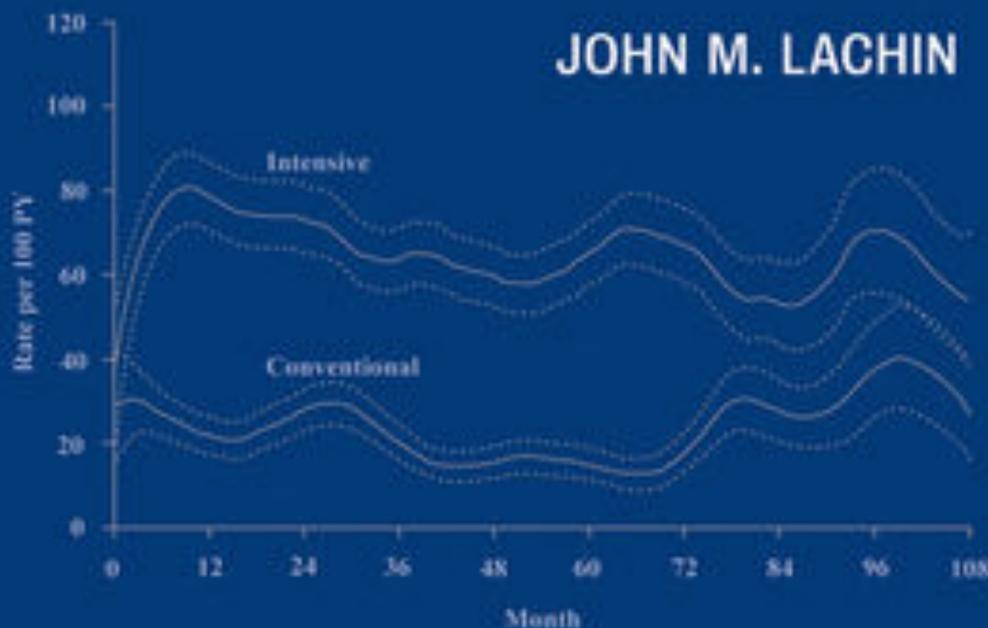


SECOND EDITION

BIOSTATISTICAL METHODS

The Assessment of Relative Risks

JOHN M. LACHIN



WILEY SERIES IN PROBABILITY AND STATISTICS

 WILEY

WWW.
WILEY.COM

Biostatistical Methods

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

Biostatistical Methods

The Assessment of Relative Risks

Second Edition

JOHN M. LACHIN



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2011 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Lachin, John M., 1942-

Biostatistical methods : the assessment of relative risks / John M. Lachin. -- 2nd ed.

p. cm. — (Wiley series in probability and statistics ; 807)

Summary: "This book focuses on the comparison, contrast, and assessment of risks on the basis of clinical investigations. It develops basic concepts as well as deriving biostatistical methods through both the application of classical mathematical statistical tools and more modern likelihood-based theories. The first half of the book presents methods for the analysis of single and multiple 2x2 tables for cross-sectional, prospective, and retrospective (case-control) sampling, with and without matching using fixed and two-stage random effects models. The text then moves on to present a more modern likelihood- or model-based approach, which includes unconditional and conditional logistic regression; the analysis of count data and the Poisson regression model; the analysis of event time data, including the proportional hazards and multiplicative intensity models; and elements of categorical data analysis (expanded in this edition). SAS subroutines are both showcased in the text and embellished online by way of a dedicated author website. The book contains a technical, but accessible appendix that presents the core mathematical statistical theory used for the development of classical and modern statistical methods"—Provided by publisher.

Includes bibliographical references and index.

ISBN 978-0-470-50822-0 (hardback)

1. Medical statistics. 2. Health risk assessment—Statistical methods. 3. Medicine—Research—Statistical methods. I. Title.

RA409.L33 2010

610.72—dc22

2010018482

Printed in the United States of America.

To my family

Contents

Preface	xix
Preface to First Edition	xxi
1 Biostatistics and Biomedical Science	1
1.1 Statistics and the Scientific Method	1
1.2 Biostatistics	2
1.3 Natural History of Disease Progression	3
1.4 Types of Biomedical Studies	5
1.5 Studies of Diabetic Nephropathy	7
2 Relative Risk Estimates and Tests for Independent Groups	13
2.1 Probability as a Measure of Risk	14
2.1.1 Prevalence and Incidence	14
2.1.2 Binomial Distribution and Large Sample Approximations	14
2.1.3 Asymmetric Confidence Limits	16
2.1.4 Case of Zero Events	19
2.2 Measures of Differential or Relative Risk	19
2.3 Large Sample Distribution	23
2.3.1 Risk Difference	23
2.3.2 Relative Risk	24

2.3.3	Odds Ratio	26
2.4	Sampling Models: Likelihoods	28
2.4.1	Unconditional Product Binomial Likelihood	28
2.4.2	Conditional Hypergeometric Likelihood	28
2.4.3	Maximum Likelihood Estimates	30
2.4.4	Asymptotically Unbiased Estimates	30
2.5	Exact Inference	32
2.5.1	Confidence Limits	32
2.5.2	Fisher-Irwin Exact Test	33
2.6	Large Sample Inferences	36
2.6.1	General Considerations	36
2.6.2	Unconditional Test	38
2.6.3	Conditional Mantel-Haenszel Test	39
2.6.4	Cochran's Test	40
2.6.5	Likelihood Ratio Test	42
2.6.6	Test-Based Confidence Limits	42
2.6.7	Continuity Correction	43
2.6.8	Establishing Equivalence or Noninferiority	45
2.7	SAS PROC FREQ	48
2.8	Other Measures of Differential Risk	52
2.8.1	Attributable Risk Fraction	52
2.8.2	Population Attributable Risk	53
2.8.3	Number Needed to Treat	56
2.9	Polychotomous and Ordinal Data	56
2.9.1	Multinomial Distribution and Large Sample Approximation	56
2.9.2	Pearson Chi-Square Test	57
2.9.3	Pearson Goodness-of-Fit Test	59
2.9.4	Logits	60
2.10	Two Independent Groups with Polychotomous Response	61
2.10.1	Large Sample Test of Proportions	61
2.10.2	The Pearson Contingency Chi-Square Test	62
2.10.3	Odds Ratios	63
2.10.4	Rank Tests: Cochran-Mantel-Haenszel Mean Scores Test	63
2.11	Multiple Independent Groups	67
2.11.1	The Pearson Test	67
2.11.2	Measures of Association	69
2.11.3	Logits	69
2.11.4	Multiple Tests	69

2.11.5 Rank and Correlation Tests	73
2.11.6 The Cochran-Armitage Test for Trend	74
2.11.7 Exact Tests	76
2.12 Problems	76
3 Sample Size, Power, and Efficiency	85
3.1 Estimation Precision	86
3.2 Power of Z-Tests	87
3.2.1 Type I and II Errors and Power	87
3.2.2 Power and Sample Size	90
3.3 Test for Two Proportions	92
3.3.1 Power of the Z-Test	93
3.3.2 Relative Risk and Odds Ratio	95
3.3.3 Equivalence	96
3.3.4 Noninferiority	98
3.4 Power of Chi-Square Tests	99
3.4.1 Noncentral Chi-Square Distribution	99
3.4.2 Pearson Chi-Square Tests	100
3.4.3 The Mean Score (Rank) Test	102
3.4.4 The Cochran-Armitage Test of Trend	104
3.5 SAS PROC POWER	106
3.5.1 Test for Two Proportions	106
3.5.2 Wilcoxon Mann-Whitney Test	107
3.6 Efficiency	108
3.6.1 Pitman Efficiency	108
3.6.2 Asymptotic Relative Efficiency	110
3.6.3 Estimation Efficiency	111
3.6.4 Stratified Versus Unstratified Analysis of Risk Differences	112
3.7 Problems	115
4 Stratified-Adjusted Analysis for Independent Groups	119
4.1 Introduction	119
4.2 Mantel-Haenszel Test and Cochran's Test	121
4.2.1 Conditional Within-Strata Analysis	121
4.2.2 Marginal Unadjusted Analysis	121
4.2.3 Mantel-Haenszel Test	123
4.2.4 Cochran's Test	125
4.3 Stratified-Adjusted Estimators	126

4.3.1	Mantel-Haenszel Estimates	126
4.3.2	Test-Based Confidence Limits	127
4.3.3	Large Sample Variance of the Log Odds Ratio	128
4.3.4	Maximum Likelihood Estimate of the Common Odds Ratio	130
4.3.5	Minimum Variance Linear Estimators	131
4.3.6	MVLE Versus Mantel-Haenszel Estimates	134
4.3.7	SAS PROC FREQ	135
4.4	Nature of Covariate Adjustment	136
4.4.1	Confounding and Effect Modification	137
4.4.2	Stratification Adjustment and Regression Adjustment	138
4.4.3	When Does Adjustment Matter?	140
4.5	Multivariate Tests of Hypotheses	145
4.5.1	Multivariate Null Hypothesis	145
4.5.2	Omnibus Test	146
4.5.3	Multiple Tests	148
4.5.4	Partitioning of the Omnibus Alternative Hypothesis	149
4.6	Tests of Homogeneity	150
4.6.1	Contrast Test of Homogeneity	151
4.6.2	Cochran's Test of Homogeneity	153
4.6.3	Zelen's Test	154
4.6.4	Breslow-Day Test for Odds Ratios	155
4.6.5	Tarone Test for Odds Ratios	156
4.7	Efficient Tests of No Partial Association	156
4.7.1	Restricted Alternative Hypothesis of Association	156
4.7.2	Radhakrishna Family of Efficient Tests of Association	158
4.8	Asymptotic Relative Efficiency of Competing Tests	163
4.8.1	Family of Tests	163
4.8.2	Asymptotic Relative Efficiency	165
4.9	Maximin-Efficient Robust Tests	169
4.9.1	Maximin Efficiency	169
4.9.2	Gastwirth Scale Robust Test	170
4.9.3	Wei-Lachin Test of Stochastic Ordering	171
4.9.4	Comparison of Weighted Tests	174
4.10	Random Effects Model	175
4.10.1	Measurement Error Model	175
4.10.2	Stratified-Adjusted Estimates from Multiple 2×2 Tables	177
4.11	Power and Sample Size for Tests of Association	183
4.11.1	Power Function of the Radhakrishna Family	184

4.11.2	Power and Sample Size for Cochran's Test	186
4.12	Polychotomous and Ordinal Data	188
4.12.1	Cochran-Mantel-Haenszel Tests	188
4.12.2	Stratified-Adjusted Estimates	189
4.12.3	Vector Test of Homogeneity	191
4.12.4	Stratified Mean Scores Estimate and Test	191
4.12.5	Stratified Cochran-Armitage Test of Trend	192
4.13	Problems	193
5	Case-Control and Matched Studies	201
5.1	Unmatched Case-Control (Retrospective) Sampling	201
5.1.1	Odds Ratio	202
5.1.2	Relative Risk	204
5.1.3	Attributable Risk	205
5.2	Matching	206
5.2.1	Frequency Matching	207
5.2.2	Matched Pairs Design: Cross-Sectional or Prospective	208
5.3	Tests of Association for Matched Pairs	211
5.3.1	Exact Test	211
5.3.2	McNemar's Large Sample Test	212
5.3.3	SAS PROC FREQ	213
5.4	Measures of Association for Matched Pairs	214
5.4.1	Conditional Odds Ratio	214
5.4.2	Confidence Limits for the Odds Ratio	215
5.4.3	Conditional Large Sample Test and Confidence Limits	217
5.4.4	Mantel-Haenszel Analysis	217
5.4.5	Relative Risk for Matched Pairs	218
5.4.6	Attributable Risk for Matched Pairs	219
5.5	Pair-Matched Retrospective Study	220
5.5.1	Conditional Odds Ratio	221
5.5.2	Relative Risks from Matched Retrospective Studies	222
5.6	Power Function of McNemar's Test	223
5.6.1	Unconditional Power Function	223
5.6.2	Conditional Power Function	224
5.6.3	Other Approaches	225
5.6.4	Matching Efficiency	226
5.7	Stratified Analysis of Pair-Matched Tables	227
5.7.1	Pair and Member Stratification	227

5.7.2	Stratified Mantel-Haenszel Analysis	228
5.7.3	MVLE	229
5.7.4	Tests of Homogeneity and Association	229
5.7.5	Random Effects Model Analysis	232
5.8	Multiple Matching: Mantel-Haenszel Analysis	232
5.9	Matched Polychotomous Data	234
5.9.1	McNemar's Test	234
5.9.2	Bowker's Test of Symmetry	234
5.9.3	Marginal Homogeneity and Quasi-symmetry	235
5.10	Kappa Index of Agreement	235
5.10.1	Duplicate Gradings, Binary Characteristic	235
5.10.2	Duplicate Gradings, Polychotomous or Ordinal Characteristic	237
5.10.3	Multiple Gradings, Intraclass Correlation	239
5.11	Problems	239
6	Applications of Maximum Likelihood and Efficient Scores	247
6.1	Binomial	247
6.2	2×2 Table: Product Binomial (Unconditionally)	249
6.2.1	MLEs And Their Asymptotic Distribution	249
6.2.2	Logit Model	250
6.2.3	Tests of Significance	254
6.3	2×2 Table, Conditionally	257
6.4	Score-Based Estimate	258
6.5	Stratified Score Analysis of Independent 2×2 Tables	260
6.5.1	Conditional Mantel-Haenszel Test and the Score Estimate	260
6.5.2	Unconditional Cochran Test as a $C(\alpha)$ Test	261
6.6	Matched Pairs	263
6.6.1	Unconditional Logit Model	263
6.6.2	Conditional Logit Model	265
6.6.3	Conditional Likelihood Ratio Test	268
6.6.4	Conditional Score Test	268
6.6.5	Matched Case-Control Study	268
6.7	Iterative Maximum Likelihood	269
6.7.1	Newton-Raphson (or Newton's Method)	269
6.7.2	Fisher Scoring (Method of Scoring)	270
6.8	Problems	275

7	Logistic Regression Models	283
7.1	Unconditional Logistic Regression Model	283
7.1.1	General Logistic Regression Model	283
7.1.2	Logistic Regression and Binomial Logit Regression	286
7.1.3	SAS Procedures	288
7.1.4	Stratified 2×2 Tables	291
7.1.5	Family of Binomial Regression Models	293
7.2	Interpretation of the Logistic Regression Model	294
7.2.1	Model Coefficients and Odds Ratios	294
7.2.2	Class Effects in PROC LOGISTIC	300
7.2.3	Partial Regression Coefficients	302
7.2.4	Model Building: Stepwise Procedures	304
7.2.5	Disproportionate Sampling	307
7.2.6	Unmatched Case-Control Study	308
7.3	Tests of Significance	309
7.3.1	Likelihood Ratio Tests	309
7.3.2	Efficient Scores Test	310
7.3.3	Wald Tests	312
7.3.4	SAS PROC GENMOD	314
7.3.5	Robust Inferences	317
7.3.6	Power and Sample Size	321
7.4	Interactions	325
7.4.1	Qualitative–Qualitative Covariate Interaction	326
7.4.2	Interactions with a Quantitative Covariate	330
7.5	Measures of the Strength of Association	333
7.5.1	Squared Error Loss	333
7.5.2	Entropy Loss	334
7.6	Conditional Logistic Regression Model for Matched Sets	337
7.6.1	Conditional Logistic Model	337
7.6.2	Matched Retrospective Study	340
7.6.3	Fitting the General Conditional Logistic Regression Model	341
7.6.4	Allowing for Clinic Effects in a Randomized Trial	341
7.6.5	Robust Inference	345
7.6.6	Explained Variation	348
7.6.7	Power and Sample Size	348
7.7	Models for Polychotomous or Ordinal Data	352
7.7.1	Multinomial Logistic Model	352
7.7.2	Proportional Odds Model	357

7.7.3	Conditional Models for Matched Sets	359
7.8	Random Effects and Mixed Models	359
7.8.1	Random Intercept Model	359
7.8.2	Random Treatment Effect	361
7.9	Models for Multivariate or Repeated Measures	363
7.9.1	<i>GEE</i> Repeated Measures Models	364
7.9.2	<i>GEE</i> Multivariate Models	368
7.9.3	Random Coefficient Models	369
7.10	Problems	370
8	Analysis of Count Data	381
8.1	Event Rates and the Homogeneous Poisson Model	382
8.1.1	Poisson Process	382
8.1.2	Doubly Homogeneous Poisson Model	382
8.1.3	Relative Risks	384
8.1.4	Violations of the Homogeneous Poisson Assumptions	388
8.2	Overdispersed Poisson Model	389
8.2.1	Two-Stage Random Effects Model	389
8.2.2	Relative Risks	392
8.2.3	Stratified-Adjusted Analyses	393
8.3	Poisson Regression Model	393
8.3.1	Homogeneous Poisson Regression Model	393
8.3.2	Explained Variation	401
8.3.3	Applications of Poisson Regression	401
8.4	Overdispersed and Robust Poisson Regression	402
8.4.1	Quasi-likelihood Overdispersed Poisson Regression	402
8.4.2	Robust Inference Using the Information Sandwich	404
8.4.3	Zeros-inflated Poisson Regression Model	407
8.5	Conditional Poisson Regression for Matched Sets	410
8.6	Negative Binomial Models	412
8.6.1	The Negative Binomial Distribution	412
8.6.2	Negative Binomial Regression Model	414
8.7	Power and Sample Size	416
8.7.1	Poisson Models	416
8.7.2	Negative Binomial Models	418
8.8	Multiple Outcomes	419
8.9	Problems	420

9	Analysis of Event-Time Data	429
9.1	Introduction to Survival Analysis	430
9.1.1	Hazard and Survival Function	430
9.1.2	Censoring at Random	431
9.1.3	Kaplan-Meier Estimator	432
9.1.4	Estimation of the Hazard Function	435
9.1.5	Comparison of Survival Probabilities for Two Groups	436
9.2	Lifetable Construction	441
9.2.1	Discrete Distributions: Actuarial Lifetable	443
9.2.2	Modified Kaplan-Meier Estimator	444
9.2.3	SAS PROC LIFETEST: Survival Estimation	446
9.3	Family of Weighted Mantel-Haenszel Tests	449
9.3.1	Weighted Mantel-Haenszel Test	449
9.3.2	Mantel-Logrank Test	450
9.3.3	Modified Wilcoxon Test	451
9.3.4	G^{ρ} Family of Tests	452
9.3.5	Measures of Association	453
9.3.6	SAS PROC LIFETEST: Tests of Significance	455
9.4	Proportional Hazards Models	456
9.4.1	Cox's Proportional Hazards Model	456
9.4.2	Stratified Models	460
9.4.3	Time-Dependent Covariates	461
9.4.4	Fitting the Model	461
9.4.5	Robust Inference	463
9.4.6	Adjustments for Tied Observations	464
9.4.7	Survival Function Estimation	468
9.4.8	Model Assumptions	469
9.4.9	Explained Variation	471
9.4.10	SAS PROC PHREG	473
9.5	Evaluation of Sample Size and Power	483
9.5.1	Exponential Survival	483
9.5.2	Cox's Proportional Hazards Model	486
9.6	Additional Models	491
9.6.1	Competing Risks	492
9.6.2	Interval Censoring	495
9.6.3	Parametric Models	496
9.6.4	Multiple Event Times	497
9.7	Analysis of Recurrent Events	499

9.7.1	Counting Process Formulation	500
9.7.2	Nelson-Aalen Estimator, Kernel Smoothed Estimator	502
9.7.3	Aalen-Gill Test Statistics	504
9.7.4	Multiplicative Intensity Model	507
9.7.5	Robust Estimation: Proportional Rate Models	511
9.7.6	Stratified Recurrence Models	512
9.8	Problems	513
Appendix Statistical Theory		535
A.1	Introduction	535
A.1.1	Notation	535
A.1.2	Matrices	536
A.1.3	Partition of Variation	537
A.2	Central Limit Theorem and the Law of Large Numbers	537
A.2.1	Univariate Case	537
A.2.2	Multivariate Case	540
A.3	Delta Method	541
A.3.1	Univariate Case	541
A.3.2	Multivariate Case	542
A.4	Slutsky's Convergence Theorem	543
A.4.1	Convergence in Distribution	543
A.4.2	Convergence in Probability	544
A.4.3	Convergence in Distribution of Transformations	544
A.5	Least Squares Estimation	546
A.5.1	Ordinary Least Squares	546
A.5.2	Gauss-Markov Theorem	548
A.5.3	Weighted Least Squares	548
A.5.4	Iteratively Reweighted Least Squares	550
A.6	Maximum Likelihood Estimation and Efficient Scores	551
A.6.1	Estimating Equation	551
A.6.2	Efficient Score	552
A.6.3	Fisher's Information Function	553
A.6.4	Cramér-Rao Inequality: Efficient Estimators	555
A.6.5	Asymptotic Distribution of the Efficient Score and the MLE	556
A.6.6	Consistency and Asymptotic Efficiency of the MLE	557
A.6.7	Estimated Information	558
A.6.8	Invariance Under Transformations	558
A.6.9	Independent But Not Identically Distributed Observations	560

A.7	Tests of Significance	561
A.7.1	Wald Tests	561
A.7.2	Likelihood Ratio Tests	563
A.7.3	Efficient Scores Test	565
A.8	Explained Variation	569
A.8.1	Squared Error Loss	570
A.8.2	Residual Variation	572
A.8.3	Negative Log-Likelihood Loss	573
A.8.4	Madalla's R_{LR}^2	573
A.9	Robust Inference	574
A.9.1	Information Sandwich	574
A.9.2	Robust Confidence Limits and Tests	579
A.10	Generalized Linear Models and Quasi-likelihood	579
A.10.1	Generalized Linear Models	580
A.10.2	Exponential Family of Models	581
A.10.3	Deviance and the Chi-Square Goodness of Fit	584
A.10.4	Quasi-likelihood	585
A.10.5	Conditional GLMs	588
A.11	Generalized Estimating Equations (GEE)	588
	References	593
	Author Index	617
	Subject Index	623

Preface

Ten years ago, almost to the day, I completed the first edition of this book. In the interim I and others have used the text as the basis for an M.S.- or Ph.D.-level course on biostatistical methods. My own three-hour course covered most of Chapters 1 to 8 and the Appendix. However, when my editor at John Wiley & Sons approached others who had also used the book for a course, I received a number of suggestions for expansion of the material, among the most prominent being the inclusion of methods for the analysis of polychotomous and ordinal data. Thus, in this second edition, throughout the text these methods are described. See the new Sections 2.9 to 2.11, 3.4, 3.5, 4.12, 5.8 to 5.10 and 7.7 in the table of contents. The evaluation of power and sample size for many of these methods has also been added to the text.

In addition, I have added a review of methods for the analysis of longitudinal repeated measures or multivariate observations, especially the family of models fit using generalized estimating equations (Sections 7.9 and A.11). I also now present an introduction to mixed models with fixed and random effects (Section 7.8).

Other additions include assessment of equivalence and noninferiority (Section 2.6.8), and sample size evaluation for such assessments (Sections 3.3.3 and 3.3.4), a discussion of adjustment for clinic effects in a multicenter study (Sections 7.6.4 and 7.8.2) and a description of negative binomial models for count data as an alternative to the Poisson model (Sections 8.6 and 8.7.2).

All of the methods are illustrated using SAS procedures from version 9.2. Readers are encouraged to review the SAS manuals that provide a more extensive review of the available options.

For the first edition, I established a website for the book that included all of the programs I used for all examples in the book, and all of the data sets used in the book.

This website has been updated to include the additional programs and data sets used herein. The book website is www.bsc.gwu.edu/jml/biostatmethods.

The first edition was replete with various typographical and computational errors and an errata was posted on the book website. As the book went through additional printings, I was able to correct most of these errors. I have been much more diligent in proofing the second edition but expect that I and others will find errors in this edition as well. Please check the website for an *Erratum* to this edition, and please bring any errors to my attention at lachin@gwu.edu.

I greatly appreciate the comments and corrections from those who read the first edition or used it for courses. I hope that this second edition will provide a book that is more useful for a broader range of curricula. While I hope that the book will be a useful technical reference, my basic objective for this edition, as for the first edition, has been to provide a graduate-level text that spans the classical and more modern spectrum of biostatistical methods. To that end I trust that the book will be useful for students, faculty, and the profession in general.

JOHN M. LACHIN

Rockville, Maryland

PREFACE TO FIRST EDITION

In 1993 to 1994 I led the effort to establish a graduate program in biostatistics at the George Washington University. The program, which I now direct, was launched in 1995 and is a joint initiative of the Department of Statistics, the Biostatistics Center (which I have directed since 1988) and the School of Public Health and Health Services. Biostatistics has long been a specialty of the statistics faculty, starting with Samuel Greenhouse, who joined the faculty in 1946. When Jerome Cornfield joined the faculty in 1972, he established a two-semester sequence in biostatistics (Statistics 225-6) as an elective for the graduate program in statistics (our 200 level being equivalent to the 600 level in other schools). Over the years these courses were taught by many faculty as a lecture course on current topics. With the establishment of the graduate program in biostatistics, however, these became pivotal courses in the graduate program and it was necessary that Statistics 225 be structured so as to provide students with a review of the foundations of biostatistics.

Thus I was faced with the question “what are the foundations of biostatistics?” In my opinion, biostatistics is set apart from other statistics specialties by its focus on the assessment of risks and relative risks through clinical research. Thus biostatistical methods are grounded in the analysis of binary and count data such as in 2×2 tables. For example, the Mantel-Haenszel procedure for stratified 2×2 tables forms the basis for many families of statistical procedures such as the G^p family of modern statistical tests in the analysis of survival data. Further, all common medical study designs, such as the randomized clinical trial and the retrospective case-control study, are rooted in the desire to assess relative risks. Thus I developed Statistics 225, and later this text, around the principle of the assessment of relative risks in clinical investigations.

In doing so, I felt that it was important first to develop basic concepts and derive core biostatistical methods through the application of classical mathematical statistical tools, and then to show that these and comparable methods may also be developed through the application of more modern, likelihood-based theories. For example, the large sample distribution of the Mantel-Haenszel test can be derived using the large sample approximation to the hypergeometric and the Central Limit Theorem, and also as an efficient score test based on a hypergeometric likelihood.

Thus the first five chapters present methods for the analysis of single and multiple 2×2 tables for cross-sectional, prospective and retrospective (case-control) sampling, without and with matching. Both fixed and random effects (two-stage) models are employed. Then, starting in Chapter 6 and proceeding through Chapter 9, a more modern likelihood or model-based treatment is presented. These chapters broaden the scope of the book to include the unconditional and conditional logistic regression models in Chapter 7, the analysis of count data and the Poisson regression model in Chapter 8, and the analysis of event time data including the proportional hazards and multiplicative intensity models in Chapter 9. Core mathematical statistical tools employed in the text are presented in the Appendix. Following each chapter problems are presented that are intended to expose the student to the key mathematical statistical derivations of the methods presented in that chapter, and to illustrate their application and interpretation.

Although the text provides a valuable reference to the principal literature, it is not intended to be exhaustive. For this purpose, readers are referred to any of the excellent existing texts on the analysis of categorical data, generalized linear models and survival analysis. Rather, this manuscript was prepared as a textbook for advanced courses in biostatistics. Thus the course (and book) material was selected on the basis of its current importance in biostatistical practice and its relevance to current methodological research and more advanced methods. For example, Cornfield's approximate procedure for confidence limits on the odds ratio, though brilliant, is no longer employed because we now have the ability to readily perform exact computations. Also, I felt it was more important that students be exposed to over-dispersion and the use of the information sandwich in model-based inference than to residual analysis in regression models. Thus each chapter must be viewed as one professor's selection of relevant and insightful topics.

In my Statistics 225 course, I cover perhaps two-thirds of the material in this text. Chapter 9, on survival analysis, has been added for completeness, as has the section in the Appendix on quasi-likelihood and the family of generalized linear models. These topics are covered in detail in other courses. My detailed syllabus for Statistics 225, listing the specific sections covered and exercises assigned, is available at the Biostatistics Center web site (www.bsc.gwu.edu/jml/biostatmethods). Also, the data sets employed in the text and problems are available at this site or the web site of John Wiley and Sons, Inc. (www.wiley.com).

Although I was not trained as a mathematical statistician, during my career I have learned much from those with whom I have been blessed with the opportunity to collaborate (chronologically): Jerry Cornfield, Sam Greenhouse, Nathan Mantel, and Max Halperin, among the founding giants in biostatistics; and also Robert Smythe, L.J. Wei, Peter Thall, K.K. Gordon Lan and Zhaohai Li, among others, who are among the best of their generation. I have also learned much from my students, who have always sought to better understand the rationale for biostatistical methods and their application.

I especially acknowledge the collaboration of Zhaohai Li, who graciously agreed to teach Statistics 225 during the fall of 1998, while I was on sabbatical leave. His detailed reading of the draft of this text identified many areas of ambiguity and greatly improved the mathematical treatment. I also thank Costas Cristophi for typing my lecture notes, and Yvonne Sparling for a careful review of the final text and programming assistance. I also wish to thank my present and former statistical collaborators at the Biostatistics Center, who together have shared a common devotion to the pursuit of good science: Raymond Bain, Oliver Bautista, Patricia Cleary, Mary Foulkes, Sarah Fowler, Tavia Gordon, Shuping Lan, James Rochon, William Rosenberger, Larry Shaw, Elizabeth Thom, Desmond Thompson, Dante Verme, Joel Verter, Elizabeth Wright, and Naji Younes, among many.

Finally, I especially wish to thank the many scientists with whom I have had the opportunity to collaborate in the conduct of medical research over the past 30 years: Dr. Joseph Schachter, who directed the Research Center in Child Psychiatry where I worked during graduate training; Dr. Leslie Schoenfeld, who directed the National Cooperative Gallstone Study; Dr. Edmund Lewis, who directed the Collaborative

Study Group in the conduct of the Study of Plasmapheresis in Lupus Nephritis and the Study of Captopril in Diabetic Nephropathy; Dr. Thomas Garvey, who directed the preparation of the New Drug Application for treatment of gallstones with ursodiol; Dr. Peter Stacpoole, who directed the Study of Dichloroacetate in the Treatment of Lactic Acidosis; and especially Drs. Oscar Crofford, Saul Genuth and David Nathan, among many others, with whom I have collaborated since 1982 in the conduct of the Diabetes Control and Complications Trial, the study of the Epidemiology of Diabetes Interventions and Complications, and the Diabetes Prevention Program. The statistical responsibility for studies of such great import has provided the dominant motivation for me to continually improve my skills as a biostatistician.

JOHN M. LACHIN

Rockville, Maryland

Biostatistics and Biomedical Science

1.1 STATISTICS AND THE SCIENTIFIC METHOD

The aim of all biomedical research is the acquisition of new information so as to expand the body of knowledge that comprises the biomedical sciences. This body of knowledge consists of three broad components:

1. descriptions of phenomena in terms of observable characteristics of elements or events;
2. descriptions of associations among phenomena;
3. descriptions of causal relationships between phenomena.

The various sciences can be distinguished by the degree to which each contains knowledge of each of these three types. The hard sciences (e.g., physics and chemistry) contain large bodies of knowledge of the third kind — causal relationships. The soft sciences (e.g., the social sciences) principally contain large bodies of information of the first and second kind — phenomenological and associative.

None of these descriptions, however, are exact. To quote the philosopher and mathematician Jacob Bronowski (1973):

All information is imperfect. We have to treat it with humility.... Errors are inextricably bound up with the nature of human knowledge....

Thus, every science consists of shared information, all of which, to some extent, is uncertain.

When a scientific investigator adds to the body of scientific knowledge, the degree of uncertainty about each piece of information is described through statistical assessments of the probability that statements are either true or false. Thus, the

language of science is statistics, for it is through the process of statistical analysis and interpretation that the investigator communicates the results to the scientific community. The syntax of this language is probability, because the laws of probability are used to assess the inherent uncertainty, errors, or precision of estimates of population parameters, and probabilistic statements are used as the basis for drawing conclusions.

The means by which the investigator attempts to control the degree of uncertainty in research conclusions is the application of the scientific method. In a nutshell, the scientific method is a set of strategies, based on common sense and statistics, that is intended to minimize the degree of uncertainty and maximize the degree of validity of the resulting knowledge. Therefore, the scientific method is deeply rooted in statistical principles.

When considered sound and likely to be free of error, such knowledge is termed scientifically valid. The designation of scientific validity, however, is purely subjective. The soundness or validity of any scientific result depends on the manner in which the observations were collected, that is, on the design and conduct of the study, as well as the manner in which the data were analyzed.

Therefore, in the effort to acquire scientifically valid information, one must consider the statistical aspects of all elements of a study:– its design, execution and analysis. To do so requires a firm understanding of the statistical basis for each type of study and for the analytic strategies commonly employed to assess a study's objectives.

1.2 BIOSTATISTICS

Biostatistics is characterized principally by the application of statistical principles to the biological/biomedical sciences, in contrast to other areas of application of statistics, such as psychometrics and econometrics. Thus, *biostatistics* refers to the development of statistical methods for, and the application of statistical principles to, the study of biological and medical phenomena.

Biomedical research activities range from the study of cellular biology to clinical therapeutics. At the basic physical level it includes *bench research*, the study of genetic, biochemical, physiological, and biological processes, such as the study of genetic defects, metabolic pathways, kinetic models, and pharmacology. Although some studies in this realm involve investigation in animals and humans (*in vivo*), many of these investigations are conducted in "test tubes" (*in vitro*). The ultimate objective of these inquiries is to advance our understanding of the pathobiology or pathophysiology of human diseases and of the potential mechanisms for their treatment.

Clinical research refers to direct observation of the clinical features of populations. This includes *epidemiology*, which can be broadly defined as the study of the distribution and etiology of human disease. Some elements, such as infectious disease epidemiology, are strongly biologically based, whereas others are more heavily dependent on empirical observations within populations. The latter include such

areas as occupational and environmental epidemiology, the study of the associations between occupational and environmental exposures with the risk of specific diseases. This type of epidemiology is often characterized as *population based* because it relies on the observation of natural samples from populations.

Ultimately, bench research or epidemiologic observation leads to advances in medical therapeutics: the development of new pharmaceuticals (drugs), devices, surgical procedures, or interventions. Such therapeutic advances are often assessed using a randomized, controlled clinical trial. Such studies evaluate the biological effectiveness of the new agent (biological efficacy), the clinical effectiveness of the therapy in practice (the *intention-to-treat* comparison), as well as the incidence of adverse effects.

The single feature that most sharply distinguishes clinical biomedical research from other forms of biological research is the propensity to assess the absolute and relative risks of various outcomes within populations. *Absolute risk* refers to the distribution of a disease, or risk factors for a disease, in a population. This risk may be expressed cross-sectionally as a simple probability, or it may be expressed longitudinally over time as a hazard function (or survival function) or an intensity process. *Relative risk* refers to a measure of the difference in risks among subsets of the population with specific characteristics, such as those exposed to a risk factor versus not exposed, or those randomly assigned to a new drug treatment versus a placebo control. The relative risk of an outcome is sometimes described as a difference in the absolute risks of the outcome, the ratio of the risks, or a ratio of the odds of the outcome.

Thus, a major part of biostatistics concerns the assessment of absolute and relative risks through epidemiological studies of various types and randomized clinical trials. This, in general, is the subject of the book. This entails the study of discrete outcomes, some of which are assessed over time. It also includes many major areas of statistics that are beyond the scope of any single book. For example, the analysis of longitudinal data is another of the various types of processes studied through biostatistics. In many studies, however, interest in a longitudinal quantitative or ordinal measure arises because of its fundamental relationship to an ultimate discrete outcome of interest. For example, longitudinal analysis of quantitative serum cholesterol levels in a population is of interest because of the strong relationship between serum lipids and the risk of cardiovascular disease. Thus, this text is devoted exclusively to the assessment of the risks of discrete characteristics or events in populations.

1.3 NATURAL HISTORY OF DISEASE PROGRESSION

Underlying virtually all clinical research is some model of our understanding of the natural history of the progression of the disease under investigation. As an example, consider the study of diabetic nephropathy (kidney disease) associated with type 1 or insulin-dependent diabetes mellitus, also known as juvenile diabetes. Diabetes is characterized by a state of metabolic dysfunction in which the subject is deficient in endogenous (self-produced) insulin. Thus, the patient must administer exogenous

Table 1.1 Stages of Progression of Diabetic Nephropathy

1. Normal: Albumin excretion rate (AER) ≤ 40 mg/24 h
 2. Microalbuminuria: $40 < \text{AER} < 300$ mg/24 h
 3. Proteinuria (overt albuminuria): $\text{AER} \geq 300$ mg/24 h
 4. Renal insufficiency: Serum creatinine > 2 mg/dL
 5. End-stage renal disease: Need for dialysis or renal transplant
 6. Mortality
-

insulin by some imperfect mechanical device, such as by multiple daily injections or a continuous subcutaneous insulin infusion (CSII) device, also called a "pump". Because of technological deficiencies with the way that insulin can be administered, it is difficult to maintain normal levels of blood glucose throughout the day, day after day. The resulting hyperglycemia leads to microvascular complications, the two most prevalent being diabetic retinopathy (disease of the retina in the eye) and diabetic nephropathy, and ultimately to cardiovascular disease.

Diabetic nephropathy is known to progress through a well-characterized sequence of disease states, characterized in Table 1.1. The earliest sign of emergent kidney disease is the leakage of small amounts of protein (albumin) into urine. The amount or rate of albumin excretion can be measured from a timed urine collection in which all the urine voided over a fixed period of time is collected. From the measurement of the urine volume and the concentration of albumin in the serum and urine at specific intervals of time, it is possible to compute the albumin excretion rate (AER) expressed as the mg/24 h of albumin excreted into the urine by the kidneys.

In the normal (nondiseased) subject, the AER is no greater than 40 mg/24 h – some would say no greater than 20 or 30 mg/24 h. The earliest sign of possible diabetic nephropathy is microalbuminuria, defined as an AER > 40 mg/24 h (but < 300 mg/24 h). As the disease progresses, the next landmark is the development of definite albuminuria, defined as an AER > 300 mg/24 h. This is often termed *overt proteinuria* because it is at this level of albumin (protein) excretion that a simple dipstick test for protein in urine will be positive. This is also the point at which nephropathy, and the biological processes that ultimately lead to destruction of the kidney, are considered well established.

To then chart the further loss of kidney function, a different measure is used: the glomerular filtration rate (GFR). The glomerulus is the cellular structure that serves as the body's filtration system. As diabetic nephropathy progresses, fewer and fewer intact glomeruli remain, so that the rate of filtration declines, starting with the leakage of protein and other elements into the urine. The GFR is difficult to measure accurately. In practice, a measure of creatinine clearance, also from a timed urine collection, or a simple measure of the creatinine concentration in serum is used to monitor disease progression. Renal insufficiency is often declared when the serum creatinine exceeds 2 mg/dL. This is followed by end-stage renal disease (ESRD), at which point the patient requires frequent dialysis or renal transplantation to prolong survival. Ultimately the patient dies from the renal insufficiency or related causes if a suitable donor kidney is not available for transplantation.

Thus, the natural history of diabetic nephropathy is described by a collection of quantitative, ordinal, and qualitative assessments. In the early stages of the disease, a study might focus entirely on quantitative measures of AER. Later, during the middle stages of the disease, this becomes problematic. For example, patients with established proteinuria may be characterized over time using a measure of GFR, but the analysis will be complicated by informatively missing observations because some patients reached ESRD or died before the scheduled completion of follow-up.

However, a study that assesses the risk of discrete outcomes, such as the incidence or prevalence of proteinuria or renal insufficiency, is less complicated by such factors and is readily interpretable by physicians. For example, if a study shows that a new drug treatment reduces the mean AER by 10 mg/24 h less than that with placebo, it is difficult to establish the clinical significance of the result. On the other hand, if the same study demonstrated a relative risk of developing proteinuria of 0.65, a 35% risk reduction with drug treatment versus placebo, the clinical significance is readily apparent to most physicians.

Therefore, we shall focus on the description of the absolute and relative risks of discrete outcomes, historically the core of biostatistics.

1.4 TYPES OF BIOMEDICAL STUDIES

Biomedical research employs various types of study designs, some of which involve formal experimentation, others not, among other characteristics. In this section the characteristics and the roles of each type of study are described briefly.

Study designs can be distinguished by three principal characteristics:

1. *Number of samples*: single versus multiple samples.
2. *Source of samples*: natural versus experimental. An experimental sample is one to which a treatment or procedure has been applied by the investigator. This may or may not involve randomization as an experimental device to assign treatments to individual patients.
3. *Time course of observation*: prospective versus retrospective versus concurrent collection of measurements and observation of responses or outcome events.

Based on these characteristics, there are basically four types of designs for biomedical studies in humans: (1) the cross-sectional study, (2) the cohort study, (3) the case-control study, and (4) the randomized experiment. A more exhaustive classification was provided by Bailar et al. (1984), but these four are the principal types. Examples of each type of study are described subsequently.

The **cross-sectional study** is a study of a single natural sample with concurrent measurement of a variety of characteristics. In the review by Bailar et al. (1984), 39% of published studies were of this type. Some notable examples are the National Health and Nutritional Examination Survey (NHANES) of the relationship between health and nutrition, and the annual Health Interview Survey of the prevalence of

various diseases in the general U.S. population. Such studies have provided important descriptions of the prevalence of disease in specified populations, of the co-occurrence of the disease and other factors (i.e., associations), and of the sensitivity and specificity of diagnostic procedures.

In a **cohort study** (25% of studies), one or more samples (cohorts) of individuals, either natural or experimental samples, are followed prospectively and subsequent status is evaluated.

A **case-control study** (5% of studies) employs multiple natural samples with retrospective measurements. A sample of cases with the disease is compared to a sample of controls without the disease with respect to the previous presence of, or exposure to, some factor.

An important characteristic of cohort and case-control studies is whether or not the study employs **matching** of pairs or sets of subjects with respect to selected covariate values. Matching is a strategy to remove bias in the comparison of groups by ensuring equality of distributions of the matching covariates employed. Matching, however, changes the sample frame or the sampling unit in the analysis from the individual subject in an unmatched study to the matched set in the matched study. Thus, matched studies require analytic procedures that are different from those more commonly applied to unmatched studies.

A **randomized, controlled clinical trial or parallel-comparative trial** (15% of studies) employs two or more parallel randomized cohorts, each of which receives only one treatment in the trial. Such studies provide a controlled assessment of a new drug, therapy, diagnostic procedure, or intervention procedure. Variations of this design include the multiple-period **crossover** design and the crossed **factorial** design. Since a clinical trial uses randomization to assign each subject to receive either the active treatment or a control (e.g., drug vs. placebo), the comparison of the groups is in expectation unbiased. However, a truly unbiased study also requires other conditions such as complete and unbiased follow-up assessments.

Each of the first three types is commonly referred to as an observational or epidemiological study, in contrast to a clinical trial. It is rare, some might say impossible, that a population-based observational study will identify a single necessary and sufficient cause for a biological effect, or a 1:1 causal relationship. Almost always, a *risk factor* is identified that has a biological effect that is associated with a change in the risk of an outcome. It is only after a preponderance of evidence is accumulated from many such studies that such a risk factor may be declared to be a *causal agent*. Such was the case with the relationship between smoking and lung cancer, and the criteria employed to declare smoking a causal agent are now widely accepted (U.S. Surgeon General, 1964, 1982).

The principal advantage of the randomized controlled trial (RCT), on the other hand, is that it can provide conclusions with respect to causal relationships because other intervening factors are controlled through randomization. Thus, the RCT provides an unbiased comparison of the effects of administering one treatment versus another on the outcome in the specified population of patients, and any differences observed can be confidently ascribed to the differences between the treatments. Therefore, the distinction between a relationship based on an observational study

and one based on a randomized experiment rests on the degree to which an observed relationship might be explained by other variables or other mechanisms.

However, in no study is there an absolute guarantee that all possible influential variables are controlled, even in a randomized, controlled experiment. Also, as the extent of knowledge about the underlying natural history of a disease expands, it becomes increasingly important to account for the known or suspected risk factors in the assessment of the effects of treatments or exposures, especially in an observational cross-sectional, cohort, or case-control study. This entails the use of an appropriate statistical model for the simultaneous influence of multiple covariates on the absolute or relative risk of important outcomes or events.

Thus, the principal objective of this book is to describe methods for the assessment of risk relationships derived from each type of study, and to consider methods to adjust or control for other factors in these assessments.

1.5 STUDIES OF DIABETIC NEPHROPATHY

To illustrate the different types of studies, we close this chapter with a review of selected studies on various aspects of diabetic nephropathy.

Cross-sectional surveys such as the National Health Interview Survey (NHIS) and the National Health and Nutrition Evaluation Survey (NHANES) indicate that approximately 16 million people in the U.S. population had some form of diabetes mellitus (Harris et al., 1987) as of around 1980. The majority had what is termed type 2 or noninsulin-dependent diabetes mellitus. Approximately 10% or 1.6 million had the more severe form, termed type 1 or insulin-dependent diabetes mellitus, for which daily insulin injections or infusions are required to sustain life. Among the most important clinical features of type 1 diabetes are the development of complications related to micro- and macrovascular abnormalities, among the most severe being diabetic nephropathy (kidney disease), which ultimately leads to end-stage renal disease (ESRD) in about a third of patients. These and other national surveys indicate that approximately 35% of all ESRD in the United States is attributed to diabetes.

As an illustration of a longitudinal observational cohort study, Deckert et al. (1978) followed a cohort of 907 Danish subjects with type 1 diabetes for many years and reported the annual incidence (proportion) of new cases of proteinuria (overt albuminuria) to appear each year. They showed that the peak incidence or greatest risk occurs approximately 15 years after the onset of diabetes. Their study also showed that over a lifetime, approximately 30% of subjects develop nephropathy whereas approximately 70% do not, suggesting that there is some mechanism that protects patients from nephropathy, possibly of a genetic nature, possibly related to the lifetime exposure to hyperglycemia, or possibly related to some environmental exposure or characteristic.

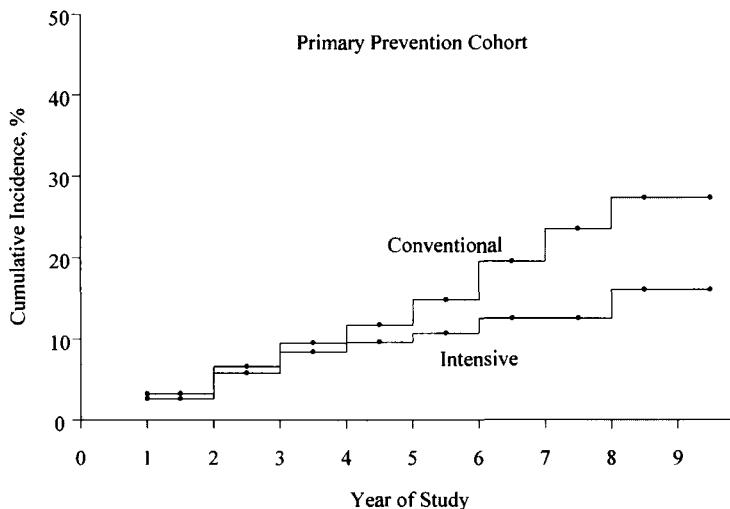
Since the discovery of insulin in the 1920s, one of the principal issues of contention in the scientific community is what was often called the *glucose hypothesis*. This hypothesis asserts that the extent of exposure to elevated levels of blood glucose

or hyperglycemia is the dominant determinant of the risk of diabetic nephropathy and other microvascular abnormalities or complications of type 1 diabetes. Among the first studies to suggest an association was a large observational study conducted by Pirart (1978a,b) in Belgium over the period 1947–1973. This study examined the association between the level of blood glucose and the prevalence (presence or absence) of nephropathy. The data were obtained from a retrospective examination of the clinical history of 4400 patients treated in a community hospital over a period of up to 25 years in some patients. The rather crude analysis consisted of figures that displayed the prevalence of nephropathy by year of diabetes duration for subgroups categorized as being in good, fair, or poor control of blood glucose levels. These figures suggest that as the mean level of hyperglycemia increases, the risk (prevalence) of nephropathy also increases. This type of study is clearly open to various types of sampling or selection biases. Nevertheless, the study provides evidence that hyperglycemia may be a strong risk factor, or is associated with the risk of diabetic nephropathy. Note that this study is not strictly a prospective cohort study because the cohort was identified later and the longitudinal observations were then obtained retrospectively.

In all of these studies, biochemical measures of renal function are used to assess the presence and extent of nephropathy. Ultimately, however, end-stage renal disease is characterized by the physiological destruction of the kidney, specifically the glomeruli, which are the cellular structures that actually perform the filtration of blood. However, the only way to determine the physical extent of glomerular damage is to conduct a morphologic evaluation of a tissue specimen obtained by a needle biopsy of the kidney. As an example of a case-control study, Chavers, Bilous, Ellis et al. (1989) conducted a retrospective study to determine whether there was an association between the presence of established nephropathy versus not (the cases vs. controls) and evidence of morphological (structural tissue) abnormalities in the kidneys (the risk factor or exposure). They showed that approximately 69% of patients with nephropathy showed morphological abnormalities versus 42% among those without nephropathy, for an odds ratio of 3.1, computed as $(0.69/0.31)$ divided by $(0.42/0.58)$. Other studies (cf. Steffes et al. (1989) show that the earliest stage of nephropathy, microalbuminuria (which they defined as an AER ≥ 20 mg/24 h), is highly predictive of progression to proteinuria, with a positive predictive value ranging from 83 to 100%. These findings established that proteinuria is indeed associated with glomerular destruction and that microalbuminuria is predictive of proteinuria. Thus, a treatment that reduces the risk of microalbuminuria can be expected to reduce the risk of progression to proteinuria, and one that reduces the risk of proteinuria will also reduce the extent of physiological damage to the kidneys.

The major question to be addressed, therefore, was whether the risk of albuminuria or nephropathy could be reduced by a treatment that consistently lowered the levels of blood glucose. By the 1980s, technological developments made an experiment (clinical trial) to test this hypothesis feasible. The level of blood glucose varies continuously over the 24-hour period, with peaks following meals and troughs before meals. It was discovered that the hemoglobin (red cells) in the blood become glycosylated when exposed to blood glucose. Thus, the percent of the total hemoglobin

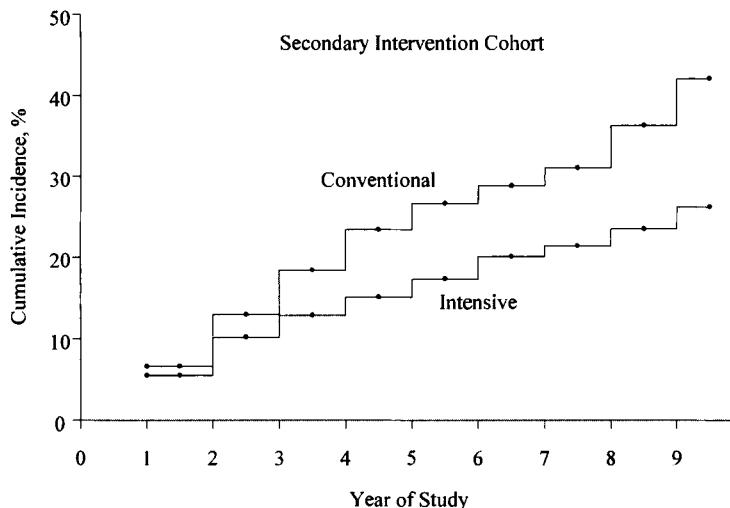
Fig. 1.1 Cumulative incidence of microalbuminuria (AER > 40 mg/24 h) over nine years of follow-up in the DCCT primary prevention cohort. Reproduced with permission.



that has become glycosylated (the HbA_{1c} %) provides an indirect measure of the mean level of hyperglycemia over the preceding four to six weeks, the half-life of the red blood cell. This made it possible to assess the average extent of hyperglycemia in individual patients. Other developments then made it possible for patients and their health care teams to control their blood sugar levels so as to lower the level of hyperglycemia, as reflected by the level of HbA_{1c} . Devices for self-monitoring of blood glucose allowed patients to measure the current level of blood glucose (mg/dL) from a drop of blood obtained by a finger prick. Patients could then alter the amount of insulin administered to keep the level of blood glucose within a desirable range. Also, a variety of types of insulin were developed, some of which acted quickly and some over long periods of time, that could be administered using multiple daily insulin injections or a pump. The health care team could then try different algorithms to vary the amount of insulin administered in response to the current level of blood glucose.

With these advances, in 1981 the National Institute of Diabetes, Digestive and Kidney Disease launched the Diabetes Control and Complications Trial (DCCT) to test the glucose hypothesis (DCCT, 1990, 1993). This was a large-scale randomized controlled clinical trial involving 1441 patients enrolled in 29 clinical centers in the United States and Canada and followed for an average of 6.5 years (4 to 9 years). Of these, the 726 patients comprising the primary prevention cohort were free of any microvascular complications ($\text{AER} \leq 40 \text{ mg/dL}$ and no retinopathy,

Fig. 1.2 Cumulative incidence of microalbuminuria (AER > 40 mg/24 h) over nine years of follow-up in the DCCT secondary intervention cohort. Reproduced with permission.



among other features); and the 715 patients comprising the secondary intervention cohort may have had minimal preexisting levels of albuminuria (AER < 200 mg/dL) and mild retinopathy. Patients were randomly assigned to receive either intensive or conventional treatment. Intensive treatment used all available means (self-monitoring four or more times a day with three or more multiple daily injections or a pump in conjunction with diet and exercise) to obtain levels of HbA_{1c} as close as possible to the normal range (< 6.05%) while attempting to avoid hypoglycemia. Hypoglycemia occurs when the blood glucose level is reduced below a physiologically safe level, resulting in dizziness and possibly coma (unconsciousness) or seizures. Conventional treatment, on the other hand, consisted of one or two daily injections of insulin with less frequent self-monitoring with the goal of maintaining the clinical well-being of the patient, but without specific glucose targets.

Figure 1.1 presents the cumulative incidence of microalbuminuria (AER > 40 mg/24 h) among the 724 patients free of microalbuminuria at baseline in the primary cohort (adapted from DCCT, 1993). The average hazard ratio for intensive versus conventional treatment (I:C) over the nine years is 0.66. This corresponds to a 34% risk reduction with intensive therapy, 95% confidence limits (2, 56%) (DCCT, 1993, 1995a). Likewise, Figure 1.2 presents the cumulative incidence of microalbuminuria among the 641 patients free of microalbuminuria at baseline in the secondary cohort. The average hazard ratio is 0.57, corresponding to a 43% (C.I.: 21, 58%) risk reduction with intensive therapy (DCCT, 1995a). These risk reductions are adjusted

for the baseline level of log AER using the proportional hazards regression model. A model that also employed a stratified adjustment for primary and secondary cohorts yields a risk reduction of 39% (21, 52%) in the combined cohorts. Similar analyses indicate a reduction of 54% (19, 74%) in the risk of overt albuminuria or proteinuria (AER > 300 mg/24 h) in the combined cohorts. Thus, intensive therapy aimed at near-normal blood glucose levels dramatically reduces the incidence of severe nephropathy, which may ultimately lead to end-stage renal disease.

However, intensive treatment was associated with an increased incidence of severe episodes of hypoglycemia (DCCT, 1993, 1995b, 1997). Over the 4770 patient years of treatment and follow-up in the intensive treatment group, 271 patients experienced 770 episodes of hypoglycemia accompanied by coma and/or seizures, or 16.3 events per 100 patient years (100 PY) of follow-up. In contrast, over the 4732 patient years in the conventional treatment group, 137 patients experienced 257 episodes, or 5.4 per 100 PY. The relative risk is 3.02 with 95% confidence limits of 2.36 to 3.86 (DCCT, 1995b, 1997). Because of substantial overdispersion of the subject-specific event rates, this confidence limit was computed using a random effects or overdispersed Poisson model.

Thus, the DCCT demonstrated that a multifaceted intensive treatment aimed at achieving near-normal levels of blood glucose greatly reduces the risk of nephropathy. The ultimate questions, however, were whether these risk reductions are caused principally by the alterations in levels of blood glucose, as opposed to changes in diet or exercise, for example, and whether there is some threshold for hyperglycemia below which there are no further reductions in risk. Thus, analyses were performed using Poisson and proportional hazards regression models, separately in the intensive and conventional treatment groups, using the current mean level of HbA_{1c} since entry into the trial as a time-dependent covariate in conjunction with numerous covariates measured at baseline. Adjusting for 25 other covariates, these models showed that the dominant determinant of the risk of proteinuria is the current level of the log mean HbA_{1c} since entry, with a 71% increase in risk per 10% increase in HbA_{1c} (such as from an HbA_{1c} of 9 to 9.9) in the conventional group, which explains approximately 5% of the variation in risk (DCCT, 1995c). Further analyses demonstrated that there is no statistical breakpoint or threshold in this risk relationship (DCCT, 1996).

These various studies and analyses, all of which concern the absolute and relative risks of discrete outcomes, show that microalbuminuria and proteinuria are associated with structural changes in renal tissue, that an intensive treatment regimen greatly reduces the risk of nephropathy, and that the principal risk factor is the lifetime exposure to hyperglycemia. Given that diabetes is the leading cause of end-stage renal disease, it can be anticipated that implementation of intensive therapy in the wide population with type 1 diabetes will ultimately reduce the progression of nephropathy to end-stage renal disease, with pursuant reductions in the incidence of morbidity and mortality caused by diabetic kidney disease and the costs to the public. The methods used to reach these conclusions, and their statistical basis, are described in the chapters to follow.

Throughout the book, selected programs are cited that perform specific computations, or access a particular data set. These and many more programs are provided

at the author's website for the book at www.bsc.gwu.edu/jml/biostatmethods. This includes all of the data sets employed in the book and the programs used for all examples.

Relative Risk Estimates and Tests for Independent Groups

The core of biostatistics relates to the evaluation and comparison of the risks of disease and other health outcomes in specific populations. Among the many different designs, the most basic is the comparison of two independent groups of subjects drawn from two different populations. This could be a cross-sectional study comparing the current health status of those with, versus those without, a specific exposure of interest; or a longitudinal cohort study of the development of health outcomes among a group of subjects exposed to a purported risk factor versus a group not so exposed; or a retrospective study comparing the previous exposure risk among independent (unmatched) samples of cases of the disease versus controls; or perhaps a clinical trial where the health outcomes of subjects are compared among those randomly assigned to receive the experimental treatment versus those assigned to receive the control treatment. Each of these cases will involve a comparison of the proportions with the response or outcome between the two groups.

Many texts provide a review of the methods for comparison of the risks or probabilities of an outcome between groups. These include the classic text by Fleiss (1981), and the updated edition of Fleiss et al. (2003), and many texts on statistical methods for epidemiology such as Breslow and Day (1980, 1987), Sahai and Khurshid (1995), Selvin (1996), and Kelsey et al. (1996), among many. Because this book is intended principally as a graduate text, readers are referred to these books for review of other topics not covered herein.

2.1 PROBABILITY AS A MEASURE OF RISK

2.1.1 Prevalence and Incidence

The simplest data structure in biomedical research is a sample of n independent and identically distributed (*i.i.d.*) Bernoulli observations $\{y_i\}$ from a sample of n subjects ($i = 1, \dots, n$) drawn at random from a population with a probability π of a characteristic of interest such as death or worsening, or perhaps survival or improvement. The character of interest is often referred to as the positive response, the outcome, or the event. Thus, Y is a binary random variable such that $y_i = I\{\text{positive response for the } i\text{th observation}\}$, where $I\{\cdot\}$ is the indicator function: $I\{\cdot\} = 1$ if true, 0 if not. The total number of subjects in the sample with a positive response is $x = \sum_i y_i$, and the simple proportion with a positive response in the sample is $p = x/n$.

The *prevalence* of a characteristic is the probability π in the population, or the proportion p in a sample, with that characteristic present in a cross section of the population at a specific point in time. For example, the prevalence of adult-onset type 2 diabetes as of 1980 was estimated to be approximately 6.8% of the U.S. population based on the National Health and Nutrition Examination Survey (NHANES) (Harris et al., 1987). Half of those who met the criteria for diabetes on an oral glucose tolerance test (3.4%) were previously undiagnosed. In such a study, n is the total sample size of whom x have the positive characteristic (diabetes).

The *incidence* of an *event* (the positive characteristic) is the probability π in the population, or the proportion p in a sample, that acquire the positive characteristic or experience an event over an interval of time among those who were free of the characteristic at baseline. In this case, n is the sample size at risk in a prospective longitudinal follow-up study, of whom x experience the event over a period of time. For example, from the annual National Health Interview Survey (NHIS) it is estimated that the incidence of a new diagnosis of diabetes among adults in the U.S. population is 2.42 new cases per 1000 in the population per year (Kenny et al., 1995).

Such estimates of the prevalence of a characteristic, or the incidence of an event, are usually simple proportions based on a sample of n *i.i.d.* Bernoulli observations.

2.1.2 Binomial Distribution and Large Sample Approximations

Whether from a cross-sectional study of prevalence or a prospective study of incidence, the number of positive responses X is distributed as binomial with probability π , or

$$P(x) = B(x; \pi, n) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad (2.1)$$

where $E(X) = n\pi$ and $V(X) = n\pi(1 - \pi)$. Since $E(X) = n\pi$, a natural moment estimate of π is p , where p is the simple proportion of events $p = x/n$. This is also the maximum likelihood estimate. From the normal approximation to the binomial,

it is well known that X is normally distributed asymptotically (in large samples) as

$$X \xrightarrow{d} \mathcal{N} [n\pi, n\pi(1 - \pi)], \quad (2.2)$$

from which

$$p \xrightarrow{d} \mathcal{N} [\pi, \pi(1 - \pi)/n]. \quad (2.3)$$

These expressions follow from the central limit theorem because x can be expressed as the n th partial sum of a potentially infinite series of *i.i.d.* random variables $\{y_i\}$ (see Section A.2). Thus, p is the mean of a set of *i.i.d.* random variables, $p = \bar{y} = \sum_i y_i/n$.

As described in the Appendix, (2.3) is a casual notation for the asymptotic distribution of p or of \bar{y} . More precisely, we would write

$$\lim_{n \rightarrow \infty} \sqrt{n}(p_n - \pi) \xrightarrow{d} \mathcal{N} [0, \pi(1 - \pi)], \quad (2.4)$$

which indicates that as the sample size becomes infinitely large, the proportion p_n converges in distribution to the normal distribution and that p is a \sqrt{n} -consistent estimator for π . In this notation, the variance is a fixed quantity, whereas in (2.3) the variance $\downarrow 0$ as $n \rightarrow \infty$.

The expression for the variance in (2.3), $V(p) = \pi(1 - \pi)/n$, is the *large sample variance* that is used in practice with finite samples to compute a test of significance or a confidence interval. A test of the null hypothesis that the probability is a specified value, $H_0: \pi = \pi_0$ against the alternative hypothesis $H_1: \pi = \pi_1 \neq \pi_0$, is then provided by a large sample test

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}, \quad (2.5)$$

where Z is asymptotically distributed as standard normal. We would then reject H_0 against the alternative H_1 , two-sided, for values $|z| \geq Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the upper two-sided normal distribution percentile at level α ; for example, for $\alpha = 0.05$, $Z_{0.975} = 1.96$.

Since $p \xrightarrow{p} \pi$, then from Slutsky's convergence theorem, (A.45) in Section A.4, a consistent estimate of the variance is $\widehat{V}(p) = p(1 - p)/n$. This yields the usual large sample confidence interval at level $1 - \alpha$ for a proportion with lower and upper confidence limits on π obtained as

$$(\widehat{\pi}_L, \widehat{\pi}_U) = p \pm Z_{1-\alpha/2} \sqrt{p(1 - p)/n}. \quad (2.6)$$

However, these confidence limits are not bounded by (0,1), meaning that for values of p close to 0 or 1, or for small sample sizes, the upper limit may exceed 1 or the lower limit be less than 0.

2.1.3 Asymmetric Confidence Limits

2.1.3.1 Exact Confidence Limits One approach that ensures that the confidence limits are bounded by $(0,1)$ is an *exact computation* under the binomial distribution, often called the Clopper-Pearson confidence limits (Clopper and Pearson, 1934). In this case the upper confidence limit π_U is the solution to the equation

$$\sum_{a=0}^x B(a; \pi, n) = \alpha/2 \quad (2.7)$$

and the lower confidence limit π_L is the solution to

$$\sum_{a=x}^n B(a; \pi, n) = \alpha/2.$$

Such confidence limits are not centered about p and thus are called *asymmetric* confidence limits.

A solution of these equations may be obtained by iterative computations. Alternatively, Clopper and Pearson show that these limits may be obtained from the relationship between the cumulative F -distribution and the incomplete beta function, of which the binomial is a special case, see, e.g., Wilks (1962). With a small value of np , confidence limits may also be obtained from the Poisson approximation to the binomial distribution. Computations of the exact limits are readily obtained using commercial software such as StatXact.

2.1.3.2 Logit Confidence Limits Another approach is to consider a function $g(\pi)$ such that the inverted confidence limits based on $g(\pi)$ are contained in the interval $(0, 1)$. One convenient function is the *logit* transformation,

$$\theta = g(\pi) = \log[\pi/(1 - \pi)], \quad (2.8)$$

where throughout, \log is the natural logarithm to the base e . The logit plays a central role in the analysis of binary (Bernoulli) data. The quantity $O = \pi/(1 - \pi)$ is the *odds* of the characteristic of interest or an event in the population, such as $O = 2$ for an odds of 2:1 when $\pi = 2/3$. The inverse logit or *logistic function*

$$\pi = g^{-1}(\theta) = \frac{e^{\theta}}{1 + e^{\theta}} = \frac{e^{\theta}}{1 + e^{\theta}} \quad (2.9)$$

then transforms the odds back to the probability.

Woolf (1955) was among the first to describe the asymptotic distribution of the log odds. Using the delta (δ)-method (see Section A.3), then asymptotically

$$E[g(\pi)] = E \left[\log \left(\frac{\pi}{1 - \pi} \right) \right] \cong \log \left(\frac{\pi}{1 - \pi} \right) = g(\pi), \quad (2.10)$$

and thus $\hat{\theta} = g(p)$ provides a consistent estimate of $\theta = g(\pi)$. The large sample variance of the estimate is

$$\begin{aligned} V(\hat{\theta}) &= V \left[\log \left(\frac{p}{1-p} \right) \right] \cong \left[\frac{d}{d\pi} \log \left(\frac{\pi}{1-\pi} \right) \right]^2 V(p) \\ &= \left(\frac{1}{\pi(1-\pi)} \right)^2 \frac{\pi(1-\pi)}{n} = \frac{1}{n\pi(1-\pi)}, \end{aligned} \quad (2.11)$$

where \cong means "asymptotically equal to." Because p is a consistent estimator of π it follows from Slutsky's theorem (A.45) that the variance can be consistently estimated by substituting p for π to yield

$$\hat{V}(\hat{\theta}) = \frac{1}{np(1-p)}. \quad (2.12)$$

Further, from another tenet of Slutsky's theorem (A.47) it follows that asymptotically

$$\hat{\theta} = \log \left(\frac{p}{1-p} \right) \stackrel{d}{\approx} \mathcal{N} \left[\log \left(\frac{\pi}{1-\pi} \right), \frac{1}{n\pi(1-\pi)} \right]. \quad (2.13)$$

Further, because $\hat{V}(\hat{\theta})$ is consistent for $V(\hat{\theta})$, it also follows from Slutsky's theorem (A.44) that asymptotically

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} = \frac{\log \left(\frac{p}{1-p} \right) - \log \left(\frac{\pi}{1-\pi} \right)}{\sqrt{\hat{V} \left[\log \left(\frac{p}{1-p} \right) \right]}} \stackrel{d}{\approx} \mathcal{N}(0, 1). \quad (2.14)$$

Thus, the symmetric $1 - \alpha$ confidence limits on the logit θ are

$$(\hat{\theta}_L, \hat{\theta}_U) = \hat{\theta} \pm Z_{1-\alpha/2} \sqrt{\frac{1}{np(1-p)}}. \quad (2.15)$$

Applying the inverse (logistic) function in (2.9) yields the asymmetric confidence limits on π :

$$(\hat{\pi}_L, \hat{\pi}_U) = \left[\frac{e^{\hat{\theta}_L}}{1 + e^{\hat{\theta}_L}}, \frac{e^{\hat{\theta}_U}}{1 + e^{\hat{\theta}_U}} \right], \quad (2.16)$$

that are bounded by $(0, 1)$.

2.1.3.3 Complementary log-log Confidence Limits Another convenient function is the *complementary log-log* transformation

$$\theta = g(\pi) = \log [-\log(\pi)], \quad (2.17)$$

that is commonly used in survival analysis. It can readily be shown (see Problem 2.1) that the $1 - \alpha$ confidence limits on $\theta = g(\pi)$ obtained from the asymptotic normal distribution of $\hat{\theta} = g(p) = \log[-\log(p)]$ are

$$(\hat{\theta}_L, \hat{\theta}_U) = \hat{\theta} \pm Z_{1-\alpha/2} \sqrt{\frac{(1-p)}{np(\log p)^2}}. \quad (2.18)$$

Applying the inverse function yields the asymmetric confidence limits on π

$$(\hat{\pi}_L, \hat{\pi}_U) = \left(\exp[-\exp(\hat{\theta}_U)], \exp[-\exp(\hat{\theta}_L)] \right) \quad (2.19)$$

that are also bounded by $(0, 1)$. Note that because the transformation includes a reciprocal, the lower limit π_L is obtained as the inverse transformation of the upper confidence limit $\hat{\theta}_U = g(\hat{\pi}_U)$ in (2.18).

2.1.3.4 Test-Inverted Confidence Limits Another set of asymmetric confidence limits was suggested by Wilson (1927) based on inverting the two-sided Z -test for a proportion in (2.5) that leads to rejection of H_0 for values $|z| \geq Z_{1-\alpha/2}$. Thus, setting $z^2 = (Z_{1-\alpha/2})^2$ yields a quadratic equation in π_0 , the roots for which provide confidence limits for π :

$$(\hat{\pi}_L, \hat{\pi}_U) = \frac{\left[\frac{Z_{1-\alpha/2}^2}{2n} + p \right] \pm \sqrt{\frac{Z_{1-\alpha/2}^2}{4n} \left[\frac{Z_{1-\alpha/2}^2}{n} + 4p(1-p) \right]}}{\frac{Z_{1-\alpha/2}^2}{n} + 1}. \quad (2.20)$$

Newcombe (1998), among others, has shown that these limits provide good coverage probabilities relative to the exact Clopper-Pearson limits. Miettinen (1976) also generalized this derivation to confidence limits for odds ratios (see Section 2.6.6).

Example 2.1 Hospital Mortality

In a particular hospital assume that $x = 2$ patients died postoperatively out of $n = 46$ patients who underwent coronary artery bypass surgery during a particular month. Then $p = 0.04348$, with $\hat{V}(p) = 0.0009041$ and estimated standard error $S.E.(p) = 0.030068$, that yields 95% large sample confidence limits from (2.6) of $(-0.01545, 0.10241)$, the lower limit value less than 0 being clearly invalid. The exact computation of (2.7) using StatXact yields limits of $(0.00531, 0.1484)$. The logit transformation $\hat{\theta} = g(p) = \log[p/(1-p)]$ yields $\hat{\theta} = \log(2/44) = -3.091$ with estimated variance $\hat{V}(\hat{\theta}) = 0.5227$ and $S.E.(\hat{\theta}) = 0.723$. From (2.15) this yields 95% confidence limits on θ of $(-4.508, -1.674)$. The logistic function of these limits yields 95% confidence limits for π of $(0.0109, 0.1579)$, that differ slightly from the exact limits. Likewise, the complementary log-log transformation $\hat{\theta} = g(p) = \log[-\log(p)]$ yields $\hat{\theta} = 1.1428$ with $S.E.(\hat{\theta}) = 0.22056$. From (2.18) this yields 95% confidence limits on θ of $(0.7105, 1.5751)$. The inverse function of

these limits yields 95% confidence limits for π of (0.00798, 0.13068), that compare favorably to the exact limits. Finally, the test-inverted confidence limits from (2.20) are (0.012005, 0.14533).

With only two events in 46 subjects, clearly the exact limits are preferred. However, even in this case, the large sample approximations are satisfactory, other than the ordinary large sample limits based on the asymptotic normal approximation to the distribution of p itself.

2.1.4 Case of Zero Events

In some cases it is important to describe the confidence limits for a probability based on a sample of n observations of which none have the positive characteristic present, or have experienced an event, such that x and p are both zero. From the expression for the binomial probability,

$$P(X = 0) = B(0; \pi, n) = (1 - \pi)^n. \quad (2.21)$$

One then desires a one-sided confidence interval of size $1 - \alpha$ of the form $(0, \hat{\pi}_U)$, where the upper confidence limit satisfies the relation

$$\hat{\pi}_U = \pi : B(0; \pi, n) = \alpha, \quad (2.22)$$

the ":" meaning "such that." Solving for π yields

$$\hat{\pi}_U = 1 - \alpha^{1/n}, \quad (2.23)$$

(see Louis, 1981).

For example, if $n = 60$, the 95% confidence interval for π when $x = 0$ is $(0, 0.0487)$. Thus, with 95% confidence we must admit the possibility that π could be as large as 0.049, or about 1 in 20. If, on the other hand, we desired an upper confidence limit of 1 in 100, such that $\hat{\pi}_U = 0.01$, then the total sample size would satisfy the expression $0.01 = 1 - 0.05^{1/n}$, that yields $n = 299$ (298.07 to be exact), (see Problem 2.3).

2.2 MEASURES OF DIFFERENTIAL OR RELATIVE RISK

The simplest design to compare two populations is to draw two independent samples of n_1 and n_2 subjects from each of the two populations and then to observe the numbers of subjects within each sample, x_1 and x_2 , who have a positive response or characteristic of interest. The resulting data can be summarized in a simple 2×2 table to describe the association between the binary independent variable representing membership in either of two independent groups ($i = 1, 2$), and a binary dependent variable (the response), where the response of primary interest is denoted as + and

Table 2.1 Measures of differential or relative risk

Type	θ	Expression	Domain	Null Value
<i>Risk difference (RD)</i>		$\pi_1 - \pi_2$	$[-1, 1]$	0
<i>Relative risk (RR)</i>		π_1/π_2	$(0, \infty)$	1
$\log RR$		$\log(\pi_1) - \log(\pi_2)$	$(-\infty, \infty)$	0
<i>Odds ratio (OR)</i>		$\frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}$	$(0, \infty)$	1
$\log OR$		$\log \left[\frac{\pi_1}{1 - \pi_1} \right] - \log \left[\frac{\pi_2}{1 - \pi_2} \right]$	$(-\infty, \infty)$	0

its complement as $-$. This 2×2 table of frequencies can be expressed as:

		Group		Group		Group	
		1	2	1	2	1	2
Response	+	x_1	x_2	n_{11}	n_{12}	$n_{1\bullet}$	$n_{\bullet 1}$
	-	$n_1 - x_1$	$n_2 - x_2$	n_{21}	n_{22}	$n_{2\bullet}$	$n_{\bullet 2}$
		n_1	n_2	$n_{\bullet 1}$	$n_{\bullet 2}$	N	N
						n_1	n_2
							N

(2.24)

where the \bullet subscript represents summation over the corresponding index for rows or columns. For the most part, we shall use the notation in the last table when it is unambiguous. Throughout, N denotes the total sample size. Within each group ($i = 1, 2$), the number of positive responses is distributed as binomial with probability π_i , from which $p_i \stackrel{d}{\sim} \mathcal{N}[\pi_i, \pi_i(1 - \pi_i)/n_i]$.

We can now define a variety of parameters to describe the differences in risk between the two populations, as shown in Table 2.1. Each measure is a function of the probabilities of the positive response in the two groups. The domain is the parameter space for that measure and the null value is the value of the parameter under the null hypothesis of no difference in risk between the two populations, $H_0: \pi_1 = \pi_2$. The *risk difference (RD)* is the simple algebraic difference between the probabilities of the positive response in the two groups with a domain of $[-1, 1]$ and with the value zero under the null hypothesis. The *relative risk (RR)* is the ratio of the probabilities in the two groups. It is also referred to as the *risk ratio*. The *odds ratio (OR)* is the ratio of the odds of the outcome of interest in the two groups. Both the relative risk and the odds ratio have a domain consisting of the positive real line and a null value of 1.0. To provide a symmetric distribution under the null hypothesis, it is customary to use the log of each.

Each of these measures can be viewed as an index of the differential or relative risk between the two groups and will reflect a departure from the null hypothesis when an association exists between group membership and the probability of response. Thus, the term *relative risk* is used to refer to a family of measures of the degree

of association between group and response, and is also used to refer to the specific measure defined as the risk ratio.

Each of these measures is of the form $\theta = G(\pi_1, \pi_2)$ for some function $G(\cdot, \cdot)$. Thus, each may be estimated by substituting the sample proportions p_i for the probabilities π_i to yield

$$\widehat{RD} = p_1 - p_2, \quad (2.25)$$

$$\widehat{RR} = p_1/p_2 = an_2/bn_1,$$

$$\widehat{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{ad}{bc}.$$

Because the p_i converge in probability to (are consistent estimates of) the probabilities π_i , then from Slutsky's convergence theorem (A.45) the resulting estimate $\widehat{\theta} = G(p_1, p_2)$ is a consistent estimate of the corresponding $\theta = G(\pi_1, \pi_2)$. These estimates, however, are not unbiased for finite samples.

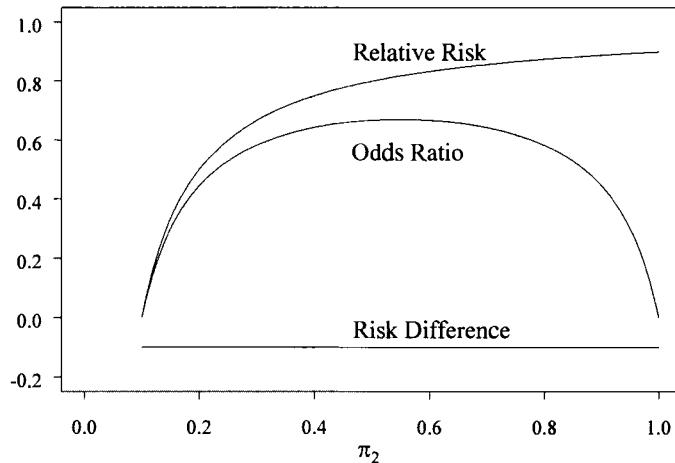
Later, in Chapter 5, we demonstrate that in a retrospective case-control study, the relative risk itself is not directly estimable without additional information. However, the odds ratio is estimable, and under the assumption that the disease (i.e., being a case) is rare, it provides an approximation to the relative risk. Thus, in some texts, the odds ratio is called the relative risk. However, the two are distinct measures of the association between group membership and the likelihood of a positive response. Later, in this chapter, we introduce other useful measures – the attributable risk and the number needed to treat.

The risk difference, relative risk, and odds ratio clearly are nonlinear functions of each other. For example, Figure 2.1 displays the values of the odds ratio and relative risk over a range of values for π_2 where the risk difference is held constant at $RD = -0.1$. As π_2 increases, the relative risk is monotonically increasing toward the null value of 1.0, indicating proportionately smaller risk reductions. For values $\pi_1 \downarrow 0$ and for values $\pi_2 \uparrow 1$, the odds ratio is $OR \cong 0$. As π_2 increases, the odds ratio increases toward the null value, reaching a maximum of 0.669 at $\pi_2 = 0.5(1 - RD) = 0.55$. The relative risk, however, continues to increase as $\pi_2 \uparrow 1$.

Thus, if we have two separate 2×2 tables, such as from two different studies, where the risk difference is the same ($RD_{(1)} = RD_{(2)}$), but where the probabilities in the control group are different [$\pi_{2(1)} \neq \pi_{2(2)}$], then the relative risks will differ, the study with the larger value of π_2 having the smaller risk reduction (relative risk closer to 1). The odds ratios will also differ if both studies have control group probabilities that are less than $\tilde{\pi}_2$ or both are greater than $\tilde{\pi}_2$. It is possible, however, that they may be approximately equal if $\pi_{2(1)} < \tilde{\pi}_2 < \pi_{2(2)}$, such as $OR = 0.66033$ for $\pi_{2(1)} = 0.46$ and $\pi_{2(2)} = 0.64$.

For any single study, the choice of the measure used to describe the study results may largely be a matter of taste or clinical preference. All three measures will reflect a difference between groups when such exists. Nonstatisticians such as physicians often find the relative risk most appealing. As we shall see, however, the odds ratio arises as the natural parameter under a conditional likelihood, and forms the basis for

Fig. 2.1 Odds ratio and relative risk over a range of values for π_2 for a fixed risk difference of -0.1 .



the analysis of case-control studies (see Chapter 5). It also forms the basis for logistic regression. For a single study with moderate or large sample size, the choice of the measure will not affect the conclusions reached. However, as we show in Chapter 4, when there are multiple 2×2 tables, such as from multiple strata, the conclusions reached may indeed depend on the measure chosen to summarize the results.

Example 2.2 Neuropathy Clinical Trial

For an example, consider a clinical trial comparing a group of subjects assigned to receive a new drug (group 1) versus a group assigned to receive placebo (2) for treatment of diabetic neuropathy where the positive response of interest is improvement in peripheral sensory perception. Patients were assigned at random to receive one of the study treatments, with 100 subjects in each group. Of these, 53 and 40 patients in each group, respectively, had a positive response. These data would be represented as

		Group				Group	
		1	2	m_1	=	1	2
Response	+	a	b	m_1	+	53	40
	-	c	d	m_2	-	47	60
		n_1	n_2	N		100	100
							200

The risk difference $\widehat{RD} = 0.53 - 0.40 = 0.13$ shows that the excess probability of improvement is 0.13. Thus, the number of patients with improved perception is increased by 13% when treated with the drug, an additional 13 patients per 100

patients so treated. The relative risk $\widehat{RR} = 0.53/0.40 = 1.325$ indicates that the total number improved is increased by 32.5% ($0.13/0.40$). Finally, the odds ratio $\widehat{OR} = (0.53 \times 0.60)/(0.40 \times 0.47) = 1.691$ indicates that the odds of improvement are increased by 69.1%, from an odds of $0.4/0.6 = 0.667$ to odds of $0.53/0.47 = 1.128$.

2.3 LARGE SAMPLE DISTRIBUTION

For the \widehat{RR} and \widehat{OR} , the domain is not symmetric about the null value. Thus, the large sample distributions are better approximated using the log transformation. In this case, since the domain encompasses the entire real line, no adjustments are required to ensure that confidence intervals are bounded within the domain of the parameter.

2.3.1 Risk Difference

The asymptotic distribution of the risk difference $\widehat{RD} = p_1 - p_2$ follows directly from that of the sample proportions themselves since the \widehat{RD} is a simple linear contrast of two independent proportions, each of which is asymptotically normally distributed. Thus,

$$p_1 - p_2 \xrightarrow{d} \mathcal{N} \left[\pi_1 - \pi_2, \frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2} \right]. \quad (2.26)$$

This is the distribution of the risk difference in general with no restrictions on the values of the probabilities π_1 and π_2 . In the context of testing the null hypothesis $H_0: \pi_1 = \pi_2$ against the general alternative hypothesis $H_1: \pi_1 \neq \pi_2$, this is termed the distribution *under the alternative hypothesis*. If we let $\theta = \pi_1 - \pi_2$ and $\widehat{\theta} = p_1 - p_2$, (2.26) is equivalent to $\widehat{\theta} \xrightarrow{d} \mathcal{N} [\theta_1, \sigma_1^2]$ under the alternative $H_1: \theta = \theta_1 \neq 0$.

Under the null hypothesis, $H_0: \pi_1 = \pi_2 = \pi$, then $\theta = \theta_0 = 0$ and the variance reduces to

$$\sigma_0^2 = \pi(1 - \pi) \left[\frac{1}{n_1} + \frac{1}{n_2} \right] = \pi(1 - \pi) \frac{N}{n_1 n_2}. \quad (2.27)$$

Thus, the distribution *under the null hypothesis* is $\widehat{\theta} \xrightarrow{d} \mathcal{N} [\theta_0, \sigma_0^2]$.

Since the p_i are consistent estimates of the π_i , then from Slutsky's theorem (A.45) the variance under the alternative can be consistently estimated by substituting the p_i for the π_i in the expression for σ_1^2 in (2.26). This yields

$$\widehat{\sigma}_1^2 = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_1}. \quad (2.28)$$

Under the null hypothesis, π is consistently estimated as

$$\widehat{\pi} = p = \frac{a + b}{n_1 + n_2} = \frac{m_1}{N}, \quad (2.29)$$

which yields a consistent estimate of the null variance σ_0^2 when substituted into (2.27).

The asymptotic distribution under the alternative leads to the usual expression for the large sample $(1 - \alpha)$ -level confidence interval for the population risk difference based on the estimate of the variance under the alternative:

$$(\hat{\theta}_L, \hat{\theta}_U) = \hat{\theta} \pm Z_{1-\alpha/2} \hat{\sigma}_{\hat{\theta}}. \quad (2.30)$$

Although common in practice, these confidence limits are not necessarily bounded by -1 and $+1$, and in rare circumstances limits are obtained that lie outside these bounds. Unlike the case of a single proportion, there is no convenient function that may be used to yield asymmetric confidence limits on the risk difference that are then bounded by $(-1, 1)$. However, by conditioning on both margins of the 2×2 table, one can obtain exact confidence limits for the risk difference that are so bounded; see Section 2.5.

The asymptotic distribution under the null hypothesis also leads to the usual expression for the large sample Z -test of the null hypothesis $H_0: \pi_1 = \pi_2$ described in Section 2.6.

Example 2.3 *Neuropathy Clinical Trial (continued)*

For the hypothetical clinical trial data presented in Example 2.2, $\hat{\theta} = \widehat{RD} = 0.13$. The large sample variance from (2.28) is estimated to be $\hat{\sigma}_{\hat{\theta}}^2 = \widehat{V}(\widehat{RD}) = (0.53 \times 0.47)/100 + (0.4 \times 0.6)/100 = 0.00489$. Thus, the $S.E.$ is $\hat{\sigma}_{\hat{\theta}} = 0.0699$, which yields a large sample 95% confidence interval of $(-0.0071, 0.2671)$.

2.3.2 Relative Risk

We now consider the distribution of the $\log \widehat{RR}$, and subsequently the $\log \widehat{OR}$, under the null and alternative hypotheses. The $\log \widehat{RR}$ is the difference in the $\log(p_i)$. The variance of the $\log(p_i)$, and thus the $V[\log(\widehat{RR})]$, is obtained using the delta (δ)-method. Likewise, the asymptotic distribution of the $\log \widehat{RR}$ is derived from that of the $\log(p_i)$ through the application of Slutsky's theorem.

In summary, by application of the δ -method, it follows (Katz et al., 1978) that for each group ($i = 1, 2$), asymptotically

$$E[\log(p_i)] \cong \log(\pi_i) \quad (2.31)$$

and that

$$V[\log(p_i)] \cong \left(\frac{d \log(\pi_i)}{d \pi_i} \right)^2 V(\pi_i) = \left(\frac{1}{\pi_i} \right)^2 \frac{\pi_i(1 - \pi_i)}{n_i} = \frac{1 - \pi_i}{n_i \pi_i}. \quad (2.32)$$

Because p_i is a consistent estimator of π_i it follows from Slutsky's theorem (A.45) that the variance can be consistently estimated by substituting p_i for π_i to yield

$$\widehat{V}[\log(p_i)] = \frac{1 - p_i}{n_i p_i}. \quad (2.33)$$

Again using Slutsky's theorem (A.47) it follows that asymptotically

$$\log(p_i) \xrightarrow{d} \mathcal{N} \left[\log(\pi_i), \frac{1 - \pi_i}{n_i \pi_i} \right], \quad (2.34)$$

and since $\widehat{V}[\log(p_i)]$ is consistent for $V[\log(p_i)]$, that

$$\frac{\log(p_i) - \log(\pi_i)}{\sqrt{\widehat{V}[\log(p_i)]}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (2.35)$$

Since $\log(\widehat{RR}) = \log(p_1) - \log(p_2)$, it follows that asymptotically

$$E[\log(\widehat{RR})] \cong \log(RR) = \log(\pi_1) - \log(\pi_2). \quad (2.36)$$

Since p_1 and p_2 are independent, then under the alternative hypothesis, with no restrictions on the values of π_1 and π_2 , the variance is

$$\sigma_1^2 \cong V[\log(\widehat{RR})] = V[\log(p_1)] + V[\log(p_2)] = \frac{1 - \pi_1}{n_1 \pi_1} + \frac{1 - \pi_2}{n_2 \pi_2}, \quad (2.37)$$

that can be consistently estimated as

$$\begin{aligned} \widehat{\sigma}_1^2 &= \widehat{V}[\log(\widehat{RR})] = \frac{1 - p_1}{n_1 p_1} + \frac{1 - p_2}{n_2 p_2} \\ &= \frac{1 - p_1}{a} + \frac{1 - p_2}{b} = \frac{1}{a} - \frac{1}{n_1} + \frac{1}{b} - \frac{1}{n_2}. \end{aligned} \quad (2.38)$$

Further, asymptotically,

$$\log(\widehat{RR}) \xrightarrow{d} \mathcal{N} \left[\log(RR), V[\log(\widehat{RR})] \right] \quad (2.39)$$

and

$$\frac{\log(\widehat{RR}) - \log(RR)}{\sqrt{\widehat{V}[\log(\widehat{RR})]}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (2.40)$$

This distribution under the alternative hypothesis is used to derive the large sample confidence limits on $\theta = \log(RR)$ as

$$(\widehat{\theta}_L, \widehat{\theta}_U) = \widehat{\theta} \pm Z_{1-\alpha/2} \widehat{\sigma}_1. \quad (2.41)$$

From these, the *asymmetric confidence limits* on the relative risk are obtained as

$$(\widehat{RR}_L, \widehat{RR}_U) = \exp \left[\widehat{\theta} \pm Z_{1-\alpha/2} \widehat{\sigma}_1 \right] = \exp (\widehat{\theta}_L, \widehat{\theta}_U) \quad (2.42)$$

that are contained within $[0, \infty)$.

Under the null hypothesis, $H_0: \pi_1 = \pi_2 = \pi$, then $\theta = \log(RR) = \theta_0 = 0$ and $\widehat{\theta} \xrightarrow{d} \mathcal{N}[\theta_0, \sigma_0^2]$. From (2.37), the variance under the null hypothesis reduces to

$$\sigma_0^2 = \frac{1 - \pi}{\pi} \left[\frac{1}{n_1} + \frac{1}{n_2} \right] = \frac{1 - \pi}{\pi} \left[\frac{N}{n_1 n_2} \right] \quad (2.43)$$

that can be consistently estimated by substituting p in (2.29) for π in (2.43).

Example 2.4 *Neuropathy Clinical Trial (continued)*

For the data in Example 2.2, the relative risk is $\widehat{RR} = 1.325$ and the $\widehat{\theta} = \log(\widehat{RR}) = \log(0.53/0.4) = 0.2814$. The large sample variance from (2.38) is estimated to be $\widehat{\sigma}_{\widehat{\theta}}^2 = \widehat{V}[\log(\widehat{RR})] = \widehat{\sigma}_1^2 = 0.47/(100 \times 0.53) + 0.6/(100 \times 0.4) = 0.02387$. The estimated *S.E.* is $\widehat{\sigma}_{\widehat{\theta}} = 0.1545$, that yields a large sample 95% confidence interval for the $\log(RR)$ of $(-0.0214, 0.5842)$. Exponentiation yields the asymmetric 95% confidence limits for the RR of $(0.9788, 1.7936)$.

Note that the relative risk herein is described as $RR_{1:2} = \pi_1/\pi_2$. If we wished to describe the relative risk as $RR_{2:1} = \pi_2/\pi_1$, we need only invert the estimate to obtain $\widehat{RR}_{2:1} = p_2/p_1 = 1/\widehat{RR}_{1:2}$. Thus, $\widehat{RR}_{2:1} = 1/1.325 = 0.755$ and $\log(\widehat{RR}_{2:1}) = \log(0.755) = -0.2814 = \log(\widehat{RR}_{1:2}^{-1}) = -\log(\widehat{RR}_{1:2})$. The estimated variance, however, is unchanged as $\widehat{V}[\log(\widehat{RR})] = 0.02387$. Thus, the confidence limits for $\log(RR_{2:1})$ are the negatives of those for $\log(RR_{1:2})$, or $(-0.5842, 0.0214)$, and the confidence limits for $RR_{2:1}$ are the reciprocals of those for $RR_{1:2}$, or $(0.5575, 1.0217) = (1.7936^{-1}, 0.9788^{-1})$.

2.3.3 Odds Ratio

The asymptotic distribution of the $\log \widehat{OR}$ may be obtained similarly to that of the $\log \widehat{RR}$, whereby the distribution of the log odds is first obtained within each group, and then the distribution of the log odds ratio is obtained as that of a linear combination of two normally distributed variates. Within each group, the log odds is simply the logit of the probability as presented in Section 2.1.3.2. In the following, however, we derive the distribution of the $\log \widehat{OR}$ through the multivariate δ -method (see Section A.3.2), starting with the asymptotic bivariate distribution of p_1 and p_2 .

Because the two groups are independent, $\mathbf{p} = (p_1 \ p_2)^T$ is asymptotically distributed as bivariate normal with mean vector $\boldsymbol{\pi} = (\pi_1 \ \pi_2)^T$ and covariance matrix under the alternative

$$\boldsymbol{\Omega}_1 = \begin{bmatrix} \frac{\pi_1(1-\pi_1)}{n_1} & 0 \\ 0 & \frac{\pi_2(1-\pi_2)}{n_2} \end{bmatrix} \quad (2.44)$$

(see Example A.2). The log odds ratio is $G(\boldsymbol{\pi}) = \log[\pi_1/(1-\pi_1)] - \log[\pi_2/(1-\pi_2)]$ with the corresponding matrix of partial derivatives (here a vector)

$$\begin{aligned} \mathbf{H}(\boldsymbol{\pi}) &= [\partial G(\boldsymbol{\pi})/\partial \pi_1 \ \partial G(\boldsymbol{\pi})/\partial \pi_2]^T \\ &= \left[\frac{1}{\pi_1(1-\pi_1)} \ \frac{-1}{\pi_2(1-\pi_2)} \right]^T. \end{aligned} \quad (2.45)$$

By application of the multivariate δ -method, the asymptotic variance of the $\log(\widehat{OR})$ under the alternative hypothesis is

$$\begin{aligned}\sigma_1^2 &= V[\log(\widehat{OR})] \cong \mathbf{H}(\boldsymbol{\pi})' \boldsymbol{\Omega}_1 \mathbf{H}(\boldsymbol{\pi}) \\ &= \frac{1}{n_1 \pi_1 (1 - \pi_1)} + \frac{1}{n_2 \pi_2 (1 - \pi_2)},\end{aligned}\quad (2.46)$$

that can be consistently estimated as

$$\begin{aligned}\widehat{\sigma}_1^2 &= \widehat{V}[\log(\widehat{OR})] = \frac{1}{n_1 p_1 (1 - p_1)} + \frac{1}{n_2 p_2 (1 - p_2)} \\ &= \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.\end{aligned}\quad (2.47)$$

This is Woolf's (1955) estimate of the variance of the log odds ratio.

From Slutsky's theorem (A.47) it follows that asymptotically

$$\log(\widehat{OR}) \xrightarrow{d} \mathcal{N}[\log(OR), \sigma_1^2] \quad (2.48)$$

and that

$$\frac{\log(\widehat{OR}) - \log(OR)}{\sqrt{\widehat{V}[\log(\widehat{OR})]}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (2.49)$$

This yields large sample confidence limits on $\theta = \log(OR)$ as

$$(\widehat{\theta}_L, \widehat{\theta}_U) = \widehat{\theta} \pm Z_{1-\alpha/2} \widehat{\sigma}_1 \quad (2.50)$$

and asymmetric confidence limits on the odds ratio

$$(\widehat{OR}_L, \widehat{OR}_U) = \exp[\widehat{\theta} \pm Z_{1-\alpha/2} \widehat{\sigma}_1] = \exp(\widehat{\theta}_L, \widehat{\theta}_U), \quad (2.51)$$

that again are contained within $[0, \infty)$.

Under the null hypothesis $H_0: \pi_1 = \pi_2 = \pi$, then $\theta = \log(OR) = \theta_0 = 0$ and $\widehat{\theta} \xrightarrow{d} \mathcal{N}[\theta_0, \sigma_0^2]$, where the null variance reduces to

$$\sigma_0^2 = \frac{1}{\pi(1 - \pi)} \left[\frac{1}{n_1} + \frac{1}{n_2} \right] = \frac{N}{\pi(1 - \pi)n_1 n_2} \quad (2.52)$$

that can be consistently estimated by substituting p in the above for π .

Example 2.5 Neuropathy Clinical Trial (continued)

Again for the data in Example 2.2, the odds ratio is $\widehat{OR} = 1.691$, and $\widehat{\theta} = \log(\widehat{OR}) = \log(1.691) = 0.526$. The estimated large sample variance from (2.47) is $\widehat{\sigma}_\theta^2 = \widehat{V}[\log(\widehat{OR})] = \widehat{\sigma}_1^2 = [1/53 + 1/47 + 1/40 + 1/60] = 0.0818$. The S.E. is $\widehat{\sigma}_\theta = 0.286$, which yields a large sample 95% confidence interval for the $\log(OR)$ of

(-0.035, 1.0862). Exponentiation yields the asymmetric 95% confidence limits for the OR of (0.9656, 2.963).

Also, as was the case for the relative risk, the odds ratio herein is defined as $OR_{1:2} = [\pi_1/(1 - \pi_1)]/[\pi_2/(1 - \pi_2)]$. If we wished to describe the odds ratio as $OR_{2:1} = [\pi_2/(1 - \pi_2)]/[\pi_1/(1 - \pi_1)]$, we would follow the same steps as described for the estimation of the inverse relative risk $RR_{2:1}$ and its confidence limits.

2.4 SAMPLING MODELS: LIKELIHOODS

In the preceding section we derived the large sample distribution of measures of relative risk starting from basic principles. Now we consider the development of the underlying likelihood based on models of sampling from a population.

2.4.1 Unconditional Product Binomial Likelihood

We can view the table of frequencies in (2.24) as arising from two independent samples of sizes n_1 and n_2 from two separate populations, of whom x_1 and x_2 are observed to have the positive characteristic of interest with probabilities π_1 and π_2 . Using the shorthand notation $a = x_1$ and $b = x_2$, the likelihood of the observations is a *product binomial likelihood*

$$\begin{aligned} L_u(\pi_1, \pi_2) &= P(a, b \mid n_1, n_2, \pi_1, \pi_2) = B(a; n_1, \pi_1)B(b; n_2, \pi_2) \quad (2.53) \\ &= \binom{n_1}{a} \pi_1^a (1 - \pi_1)^{n_1-a} \binom{n_2}{b} \pi_2^b (1 - \pi_2)^{n_2-b}. \end{aligned}$$

This is also termed the *unconditional likelihood*.

This likelihood applies to any study in which independent samples are drawn from two populations. In some cases, however, as in a cross-sectional study, one draws a single sample of size N from the population, and the sample is then cross-classified with respect to a binary independent variable that forms the two groups, and with respect to a binary dependent variable that forms the positive or negative responses. In this case the likelihood is a multinomial (quadrinomial) with four cells. However, if we condition on the group margin totals (the n_1 and n_2), then the product binomial likelihood results.

2.4.2 Conditional Hypergeometric Likelihood

The unconditional likelihood (2.53) can also be expressed in terms of the total number of positive responses m_1 since $b = m_1 - a$. This yields

$$\begin{aligned}
P(a, m_1 | n_1, n_2, \pi_1, \pi_2) & \quad (2.54) \\
&= \binom{n_1}{a} \binom{n_2}{m_1 - a} \left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \right)^a (1-\pi_1)^{n_1} \pi_2^{m_1} (1-\pi_2)^{n_2-m_1} \\
&= \binom{n_1}{a} \binom{n_2}{m_1 - a} \varphi^a (1-\pi_1)^{n_1} \pi_2^{m_1} (1-\pi_2)^{n_2-m_1}
\end{aligned}$$

where

$$\varphi = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \quad (2.55)$$

is the odds ratio.

In general, a conditional distribution is obtained as

$$f(x|y) = f(x, y)/f(y) = f(x, y)/[\int f(x, y) dx]. \quad (2.56)$$

Thus, we can obtain the conditional likelihood, conditioning on m_1 being fixed, as

$$L_c(\varphi) = \frac{P(a, m_1 | n_1, n_2, \pi_1, \pi_2)}{P(m_1 | n_1, n_2, \pi_1, \pi_2)}, \quad (2.57)$$

where

$$P(m_1 | n_1, n_2, \pi_1, \pi_2) = \sum_{a=a_\ell}^{a_u} P(a, m_1 | n_1, n_2, \pi_1, \pi_2) \quad (2.58)$$

is the summation over all possible values of the index frequency $a_\ell \leq a \leq a_u$ given the row and column margins.

It is clear that the maximum possible value of a given the fixed margins is $a_u = \min(n_1, m_1)$. Likewise, the maximum value of b is $b_u = \min(n_2, m_1)$. Since $a = m_1 - b$, the minimum possible value of a is $a_\ell = m_1 - b_u$. Thus, the limits of the summation over a are

$$\begin{aligned}
a_\ell &= \max(0, m_1 - n_2) \\
a_u &= \min(m_1, n_1).
\end{aligned} \quad (2.59)$$

Then, substituting (2.54) and (2.58) into (2.57) yields the *conditional noncentral hypergeometric likelihood*,

$$L_c(\varphi) = P(a | n_1, m_1, N, \varphi) = \frac{\binom{n_1}{a} \binom{N-n_1}{m_1-a} \varphi^a}{\sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{N-n_1}{m_1-i} \varphi^i}, \quad (2.60)$$

that is a function only of the *noncentrality parameter*, the odds ratio φ .

2.4.3 Maximum Likelihood Estimates

Because the product binomial likelihood (2.53) is the product of two independent binomial likelihoods it then follows that the maximum likelihood estimate (*MLE*) of $\pi = (\pi_1, \pi_2)$ is $\mathbf{p} = (p_1, p_2)$, and that the covariance matrix is the diagonal matrix in (2.44). From the invariance principle (Section A.6.8), the *MLEs* of the relative risk and the odds ratio, or the logs thereof, based on the unconditional likelihood are the corresponding functions of p_1 and p_2 with large sample variances as obtained above using the δ -method.

Through the conditional hypergeometric distribution in (2.60), one can also derive the maximum likelihood estimating equation for φ , the odds ratio. However, the equation does not provide a closed-form expression for the *MLE* $\hat{\varphi}$, and thus an iterative procedure such as Newton-Raphson is required (see Birch, 1964). Because this estimate of the odds ratio is based on the conditional hypergeometric likelihood, it is sometimes called the conditional *MLE* of the odds ratio to distinguish it from the unconditional *MLE* = ad/bc . These estimators are described in Chapter 6.

2.4.4 Asymptotically Unbiased Estimates

Because of the discreteness of the binomial distribution from which the frequencies are sampled, the sample estimates of the risk difference, relative risk, and odds ratio are biased. For example, consider the sample estimate of the odds ratio $\hat{\varphi} = \widehat{OR}$. From (2.60),

$$E(\hat{\varphi}) = \sum_{a=a_\ell}^{a_u} \frac{a(n_2 - m_1 + a)}{(n_1 - a)(m_1 - a)} P(a \mid n_1, m_1, N, \varphi), \quad (2.61)$$

which is undefined since $\hat{\varphi}$ is undefined for those tables where $a = a_\ell$ or $a = a_u$ such that one of the cell frequencies is zero. Alternatively, one could evaluate $E(a)$, from which one could obtain an unbiased moment estimate of φ . However, no closed-form solution exists. Similar results also apply to the expected value of the log odds ratio, for which no simple bias corrected moment estimator exists. However, simple bias corrections have been suggested.

The bias of various estimators, say $\hat{\theta}(x)$, of the logit $\log[\pi/(1 - \pi)]$ and of $\log(\pi)$ for a single binomial has been evaluated using

$$E \left[\hat{\theta}(x) \right] = \sum_{x=0}^n \hat{\theta}(x) B(x; n, \pi) \quad (2.62)$$

for a fixed value of π . Gart and Zweifel (1967) present a review of the properties of various estimators. For estimation of the logit, or the odds ratio, the most common suggestion is to add 1/2 to each cell. The resulting bias-corrected estimate of the logit $\theta = \log[\pi/(1 - \pi)]$ is

$$\hat{\theta} = \log \left[\frac{x + \frac{1}{2}}{n - x + \frac{1}{2}} \right], \quad (2.63)$$

which is unbiased except for terms of $O(n^{-2})$, or of order $1/n^2$. From the δ -method (see Section A.3), the simple logit $\log[p/(1-p)]$ is unbiased except for the remainder that is $O(n^{-1})$ or of order $1/n$. Thus, the estimator in (2.63) converges to the true value $\theta = \log[\pi/(1-\pi)]$ at a faster rate than does the simple logit. The estimated large sample variance of this estimate is

$$\widehat{V}(\widehat{\theta}) = \frac{1}{x + \frac{1}{2}} + \frac{1}{n - x + \frac{1}{2}}, \quad (2.64)$$

which is unbiased for the true asymptotic variance to $O(n^{-3})$. For computations over a range of values of the parameters, Gart and Zweifel (1967) also showed that the bias of $\widehat{\theta}$ is <1% when $np > 2$.

These results suggest that the following estimate of θ = the log odds ratio, widely known as the *Haldane-Anscombe estimate* (Haldane, 1956; Anscombe, 1956), is nearly unbiased:

$$\begin{aligned} \widehat{\theta} &= \log(\widehat{OR}) = \log \left[\frac{x_1 + \frac{1}{2}}{n_1 - x_1 + \frac{1}{2}} \right] - \log \left[\frac{x_2 + \frac{1}{2}}{n_2 - x_2 + \frac{1}{2}} \right] \\ &= \log \left[\frac{a + \frac{1}{2}}{c + \frac{1}{2}} \right] - \log \left[\frac{b + \frac{1}{2}}{d + \frac{1}{2}} \right] \end{aligned} \quad (2.65)$$

with a nearly unbiased estimate of the large sample variance most easily expressed as a modified Woolf's estimate:

$$\widehat{V}(\widehat{\theta}) = \frac{1}{a + \frac{1}{2}} + \frac{1}{b + \frac{1}{2}} + \frac{1}{c + \frac{1}{2}} + \frac{1}{d + \frac{1}{2}}. \quad (2.66)$$

Similarly, Walter (1975) showed that a nearly unbiased estimator of $\theta = \log(\pi)$ is provided by

$$\widehat{\theta} = \log(\widehat{\pi}) = \log \left[\frac{x + \frac{1}{2}}{n + \frac{1}{2}} \right] \quad (2.67)$$

and Pettigrew et al. (1986) showed that the estimated large sample variance

$$\widehat{V}(\widehat{\theta}) = \frac{1}{x + \frac{1}{2}} - \frac{1}{n + \frac{1}{2}} \quad (2.68)$$

is also nearly unbiased. Thus, the following estimate of θ = the log relative risk is nearly unbiased

$$\widehat{\theta} = \log(\widehat{RR}) = \log \left[\frac{x_1 + \frac{1}{2}}{n_1 + \frac{1}{2}} \right] - \log \left[\frac{x_2 + \frac{1}{2}}{n_2 + \frac{1}{2}} \right] \quad (2.69)$$

with a nearly unbiased estimate of the large sample variance:

$$\widehat{V}(\widehat{\theta}) = \frac{1}{x_1 + \frac{1}{2}} - \frac{1}{n_1 + \frac{1}{2}} + \frac{1}{x_2 + \frac{1}{2}} - \frac{1}{n_2 + \frac{1}{2}}. \quad (2.70)$$

These estimates can then be used to compute less biased confidence interval estimates for the parameters on each scale. Clearly, their principal advantage is with small samples. In such cases, however, exact confidence limits are readily calculated using statistical software such as StatXact.

Example 2.6 *Neuropathy Clinical Trial (continued)*

For the simple data table in Example 2.2, the bias-corrected or Haldane-Anscombe estimate of the odds ratio is $(53.5 \times 60.5)/(40.5 \times 47.5) = 1.68$, yielding a log odds ratio of 0.520 with an estimated modified Woolf's variance of 0.08096. All three quantities are slightly less than those without the added 1/2 bias correction (1.69, 0.525, and 0.0818, respectively). Likewise, the bias-corrected estimate of the log relative risk is 0.278 with an estimated variance of 0.0235, nearly equivalent to the unadjusted values of 0.28 and 0.0239, respectively. Thus, these adjustments are principally advantageous with small sample sizes.

2.5 EXACT INFERENCE

2.5.1 Confidence Limits

From the noncentral hypergeometric conditional likelihood (2.60), by recursive calculations one can determine the exact $(1 - \alpha)$ -level confidence limits $(\hat{\varphi}_L, \hat{\varphi}_U)$ on the odds ratio φ . Using the limits of summation for the index cell (a_ℓ, a_u) , the lower one-sided confidence limit at level α is that value $\hat{\varphi}_L$ which satisfies the equation

$$\alpha = \sum_{x=a}^{a_u} P(x \mid n_1, m_1, N, \hat{\varphi}_L). \quad (2.71)$$

Likewise, the upper one-sided confidence limit at level α satisfies the equation

$$\alpha = \sum_{x=a_\ell}^a P(x \mid n_1, m_1, N, \hat{\varphi}_U) \quad (2.72)$$

(Cornfield, 1956). For two-sided limits, $\alpha/2$ is employed in the above for each limit. These limits are exact in the sense that the coverage probability is at least $1 - \alpha$. Either limit can be readily computed using the recursive secant method (cf. Thisted, 1988) with the usual asymmetric large sample confidence limits from (2.51) as the starting values. When $a = a_u$ there is no solution and the lower confidence limit is 0. Likewise, when $a = a_\ell$, the upper limit is ∞ . An example is described below.

Before the advent of modern computing, the calculation of these exact confidence limits was tedious. In an important paper, Cornfield (1956) derived a simple, recursive approximate solution for these exact limits. Gart (1971) presented a slight modification of Cornfield's method that reduces the number of iterations required to reach the solution. Fleiss (1979), among others, has shown that Cornfield's method is often the most accurate procedure, relative to the exact limits, among the various approximations available, including the noniterative large sample confidence limits presented earlier. However, with the widespread availability of StatXact and other programs that readily perform the exact computations, today there is little use of Cornfield's procedure.

Since the exact confidence limits for the odds ratio are obtained by conditioning on both margins as fixed, these also yield exact confidence limits on the relative risk and the risk difference as described by Thomas and Gart (1977). For example, for

a given odds ratio φ and fixed values of n_1 , m_1 , and N , it can be shown that the odds ratio of the expected frequencies is a quadratic function in π_1 . The solution for π_1 is the root bounded by $(0, 1)$, from which the value of π_2 is obtained by subtraction (see Problem 2.7). Thus, given the upper exact confidence limit on the odds ratio $\widehat{\varphi}_U$, there are corresponding probabilities $\widehat{\pi}_{U1}$ and $\widehat{\pi}_{U2}$ for which the odds ratio of the expected frequencies equals $\widehat{\varphi}_U$. Then the corresponding exact upper limit on the risk difference is $\widehat{RD}_U = \widehat{\pi}_{U1} - \widehat{\pi}_{U2}$, and that for the relative risk is $\widehat{RR}_U = \widehat{\pi}_{U1}/\widehat{\pi}_{U2}$. Likewise, exact lower confidence limits for the probabilities $\widehat{\pi}_{L1}$ and $\widehat{\pi}_{L2}$ are obtained from the lower limit on odds ratio $\widehat{\varphi}_L$, from which the lower limits \widehat{RD}_L and \widehat{RR}_L are obtained.

The Thomas-Gart approach, however, has been criticized and alternative approaches proposed wherein the confidence limits for the risk difference or risk ratio are based on an exact distribution for which each is the corresponding noncentrality parameter. These limits can differ substantially from the Thomas-Gart limits because the corresponding exact test on which they are based is some variation of an exact test based on the product binomial likelihood originally due to Barnard (1945), that is different from the Fisher-Irwin test described below. Since the product binomial likelihood involves a nuisance parameter (π_2) in addition to the risk difference or risk ratio, where $\pi_1 = \theta + \pi_2$ or $\pi_1 = \theta\pi_2$, respectively, then the exact test involves a maximization over the value of the nuisance parameter, π_2 . Despite the fact that Barnard later retracted his test, exact confidence limits based on his procedure have been implemented in StatXact. The advantage of the Thomas-Gart confidence limits, however, is that they agree exactly with the Fisher-Irwin test, which is the most widely accepted exact test for 2×2 tables.

2.5.2 Fisher-Irwin Exact Test

Fisher (1935) and Irwin (1935) independently described an exact statistical test of the hypothesis of no association between the treatment or exposure group and the probability of the positive characteristic or response. This is expressed by the null hypothesis $H_0: \pi_1 = \pi_2$ that is equivalent to $H_0: \varphi = \varphi_0 = 1$, where φ_0 represents the odds ratio under the null hypothesis. In this case, the conditional likelihood from (2.60) reduces to the *central hypergeometric distribution*:

$$L_c(\varphi_0) = P(a \mid n_1, m_1, N, \varphi_0) = \frac{\binom{n_1}{a} \binom{N - n_1}{m_1 - a}}{\sum_{i=a}^{a_u} \binom{n_1}{i} \binom{N - n_1}{m_1 - i}}. \quad (2.73)$$

As shown in Problem 2.6, the denominator equals $\binom{N}{m_1}$ so that the conditional likelihood reduces to

$$L_c(\varphi_0) = \frac{\binom{n_1}{a} \binom{n_2}{b}}{\binom{N}{m_1}} = \frac{\binom{m_1}{a} \binom{m_2}{c}}{\binom{N}{n_1}} = \frac{n_1! n_2! m_1! m_2!}{N! a! b! c! d!}. \quad (2.74)$$

Thus, the probability of the observed table can be considered to arise from a collection of N subjects of whom m_1 have a positive response, with a of these being drawn from the n_1 subjects in group 1 and b from among the n_2 subjects in group 2 ($a + b = m_1$; $n_1 + n_2 = N$). The probability can also be considered to arise by selecting n_1 of N subjects to be members of group 1, with a of these drawn from among the m_1 with a positive response and c from among the m_2 with a negative response ($a + c = n_1$; $m_1 + m_2 = N$).

For a test of H_0 against the one-sided left or lower-tailed alternative hypothesis $H_{1<}: \pi_1 < \pi_2$, the exact one-sided left-tailed p -value is

$$p_L = \sum_{x=a}^a P(x \mid n_1, m_1, N, \varphi_0). \quad (2.75)$$

Likewise, for a test of H_0 against the one-sided right or upper-tailed alternative $H_{1>}: \pi_1 > \pi_2$, the exact one-sided right-tailed p -value is

$$p_U = \sum_{x=a}^{a_u} P(x \mid n_1, m_1, N, \varphi_0). \quad (2.76)$$

The computation of the exact p -value for a test of H_0 against the two-sided alternative, $H_{1\neq}: \pi_1 \neq \pi_2$, requires that consideration be given to the sample space under such an alternative. Among the various approaches to this problem that have been suggested, the most widely accepted is based on consideration of the total probability of all tables with an individual probability no greater than that for the observed table. To simplify notation, let $P_c(x)$ refer to the central hypergeometric probability in (2.73) for a 2×2 table with index cell frequency x . Then, the two-sided p -value is computed as

$$p = \sum_{x=a}^{a_u} I\{P_c(x) \leq P_c(a)\} P_c(x), \quad (2.77)$$

where $I\{\cdot\}$ is the indicator function defined in Section 2.1.1. Note that this will not equal double the one-tailed upper p -value, or double the lower p -value, or the addition of the two. These and other exact calculations are readily performed by SAS PROC FREQ and by StatXact, among many available programs.

The Fisher-Irwin exact test has been criticized as being too conservative because other *unconditional* tests such as that originally proposed by Barnard (1945) have been shown to yield a smaller p -value and thus are more powerful. The two-sided exact p -value for Barnard's test is defined as

$$p = \max_{\pi} \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} I\left[\left|\frac{x_1}{n_1} - \frac{x_2}{n_2}\right| \geq |p_1 - p_2|\right] P_u[x_1, x_2 \mid \pi], \quad (2.78)$$

where $P_u[x_1, x_2 \mid \pi] = L_u(\pi_1, \pi_2 \mid \pi_1 = \pi_2 = \pi)$ is the unconditional probability of each 2×2 table in (2.53). This test deals with the nuisance parameter π by computing the p -value using that value π for which the p -value is maximized. The principal reason that this test is more powerful than the Fisher-Irwin exact test is because the exact p -value includes the probability of the observed table. Because the conditional hypergeometric probability of the observed table in (2.74) is generally

greater than the product binomial unconditional probability in (2.78) for the observed table, then the contribution to the Fisher-Irwin exact test p -value is greater than that to the unconditional p -value.

Such tests, however, consider all possible 2×2 tables with n_1 and n_2 individuals in each group, and with the total number of successes m_1 ranging from 0 to $\min(n_1, n_2)$. Barnard (1949) later retracted his proposal because the set of all possible 2×2 tables includes specific tables that are clearly irrelevant. Nevertheless, many have continued to advocate his approach and his approach has been generalized. Many of these procedures are implemented in StatXact (see Mehta and Patel, 1999).

Example 2.7 *Exact Inference*

For example, consider the following 2×2 table with a small total sample size $N = 29$

		Group		15	14	29	(2.79)
		1	2				
Response	+	7	8				
	-	12	2				
		19	10				

where $p_1 = 0.37$ (7/19) and $p_2 = 0.8$ (8/10). The sample odds ratio is $\widehat{OR} = 0.1458$. The iterative conditional MLE of the odds ratio is $\widehat{\varphi} = 0.1566$, as provided by StatXact (see Chapter 6).

To compute the exact confidence limits, or conduct an exact test, we then consider the set of possible tables with values $a \in [a_\ell, a_u]$, where from (2.59) the range of possible values for a is from $a_\ell = 5$ to $a_u = 15$. A one-sided exact test of H_0 versus the one-sided alternative $H_{1<}: \pi_1 < \pi_2$ is conducted by evaluating the probability associated with all tables for which $p_1 \leq 0.37$ (and $p_2 > 0.8$). For this example, this corresponds to the probability of the set of tables for which the index frequency is $a \leq 7$. These are the tables

<table border="1"><tr><td>5</td><td>10</td></tr><tr><td>14</td><td>0</td></tr><tr><td>19</td><td>10</td></tr></table>	5	10	14	0	19	10	15	or	<table border="1"><tr><td>6</td><td>9</td></tr><tr><td>13</td><td>1</td></tr><tr><td>19</td><td>10</td></tr></table>	6	9	13	1	19	10	15	or	<table border="1"><tr><td>7</td><td>8</td></tr><tr><td>12</td><td>2</td></tr><tr><td>19</td><td>10</td></tr></table>	7	8	12	2	19	10	15	(2.80)
5	10																									
14	0																									
19	10																									
6	9																									
13	1																									
19	10																									
7	8																									
12	2																									
19	10																									
	14			14			14																			

with corresponding probabilities 0.00015, 0.00350, and 0.02924. Note that the lower limit for the index cell is $a_\ell = 5$, which is determined by the margins of the table. Thus, the lower-tailed p -value is $p_L = \sum_{a=5}^7 P(x \mid 19, 15, 29, \varphi_0) = 0.03289$. This would lead to rejection of H_0 in favor of $H_{1<}$ at the $\alpha = 0.05$ level, one-sided. Evaluating (2.72) recursively at the one-sided 0.05 level yields an exact 95% one-sided confidence interval $0 < \varphi < 0.862$.

Conversely, a one-sided exact test against the one-sided alternative $H_{1>}: \pi_1 > \pi_2$ in the opposite tail is based on the probability associated with all tables for which $p_1 \geq 0.37$. For this example, this yields the upper-tailed p -value $p_U = \sum_{x=7}^{15} P(x \mid 19, 15, 29, \varphi_0) = 0.993$. This would fail to lead to rejection of H_0 in favor of $H_{1>}$ at the $\alpha = 0.05$ level, one-sided. Evaluating (2.71) recursively at the 0.05 level (using $1 - \alpha = 0.95$) yields an exact 95% one-sided confidence interval $0.0190 < \varphi < \infty$.

A two-sided test against the two-sided alternative $H_{1\neq}: \pi_1 \neq \pi_2$ is based on the total probability of all tables with an individual probability no greater than that for the observed table, which in this example is $P(7 | 19, 15, 29, \varphi_0) = 0.02924$. Examination of the table probabilities for possible values of the index cell in the opposite tail (upper in this case) yields 0.00005 for $a = 15$, 0.0015 for $a = 14$, 0.01574 for $a = 13$, and 0.078 for $a = 12$, the last of which exceeds the probability for the observed table. Thus, for this example, the two-sided p -value equals

$$p = \sum_{x=5}^7 P(x | 19, 15, 29, \varphi_0) + \sum_{x=13}^{15} P(x | 19, 15, 29, \varphi_0) = 0.0502 \quad (2.81)$$

for which we would fail to reject H_0 at the 0.05 level.

Evaluating (2.71) and (2.72) at $\alpha = 0.025$ yields exact 95% two-sided confidence limits for the odds ratio of (0.0127, 1.0952). From these values, the Thomas and Gart (1977) exact confidence limits (-0.6885, 0.0227) are obtained for the risk difference and (0.2890, 1.0452) for the relative risk (see Problem 2.7). These limits agree with the two-sided p -value of $p \leq 0.0502$.

As described in Section 2.5.1, the program StatXact takes a different approach to constructing confidence limits for the risk difference and the relative risk, in each case using an unconditional exact test. For the risk difference, in this example StatXact provides an exact p -value based on Barnard's test of $p \leq 0.0353$ with 95% exact confidence limits of (-0.7895, -0.0299). The p -value is smaller than that from Fisher's test, and thus the confidence limits disagree with the Thomas-Gart limits derived from the Fisher-Irwin test. For the relative risk (risk ratio), StatXact computes an exact p -value of $p \leq 0.2299$ with 95% exact confidence limits of (0.1168, 1.4316) that differ substantially from the p -value and the confidence limits from the Fisher-Irwin test. The precise methods employed by StatXact are described in Mehta and Patel (1999).

2.6 LARGE SAMPLE INFERENCES

2.6.1 General Considerations

With large samples, a variety of approaches can be employed to yield a test of the null hypothesis $H_0: \pi_1 = \pi_2$ based on the asymptotic distribution of a test criterion. Since $H_0: \pi_1 - \pi_2 = 0$ implies, and is implied by, the null hypothesis $H_0: \theta = [g(\pi_1) - g(\pi_2)] = 0$ for any smooth function $g(\pi)$, then a family of tests based on any family of such smooth functions will yield asymptotically equivalent results (see Problem 2.9). Note that such a family includes the log relative risk and log odds ratio, but not the relative risk and odds ratio themselves; the log relative risk and log odds ratio scales being preferred because the null distributions are then symmetric about $\theta = \theta_0 = 0$. In this section, because all such tests can be shown to be asymptotically equivalent, we only consider the test based on the risk difference $\widehat{RD} = p_1 - p_2$, which is the basis for the usual large sample test for two proportions.

Under the Neyman-Pearson construction, one first determines a priori the significance level α to be employed, that is, the probability of a Type I false positive

error of falsely rejecting the tested or null hypothesis $H_0: \theta = 0$ when, in fact, it is true. One then specifies the nature of the alternative hypothesis of interest. This can either be a one-sided left-tailed alternative $H_{1<}: \theta < 0$, a one-sided right-tailed alternative $H_{1>}: \theta > 0$, or a two-sided alternative $H_{1\neq}: \theta \neq 0$. Each alternative then implies a different rejection region for the statistical test. For a one-sided left-tailed alternative, the rejection region consists of the lower left area of size α under the probability distribution of the test statistic under the null hypothesis. For a one-sided right-tailed alternative the rejection region consists of the upper area of size α under the null hypothesis distribution. For a two-sided test the rejection region consists of the upper and lower tail areas of size $\alpha/2$ in each tail under the null hypothesis. Although one-sided tests may be justifiable in some situations, the two-sided test is more widely used. Also, some tests are inherently two-sided. If the observed value of the test statistic falls in the rejection region for the specified alternative hypothesis, the null hypothesis is rejected with type I error probability α .

For example, consider that we wish to test $H_0: \theta = \theta_0$ for some parameter θ . Let T be a test statistic that is asymptotically normally distributed and consistent for θ , with large sample variance $\sigma_T^2(\theta)$ that may depend on θ , which is the case for proportions. For a two-sided test of H_0 versus $H_1: \theta \neq \theta_0$, the rejection region consists of all values $|T| \geq T_\alpha$, where $T_\alpha = Z_{1-\alpha/2}\widehat{\sigma}_T(\theta_0)$. Thus, the test can also be based on the standardized normal deviate, or the Z -test

$$Z = \frac{T - \theta_0}{\widehat{\sigma}_T(\theta_0)}, \quad (2.82)$$

where H_0 is rejected in favor of H_1 , two-sided, when $|z| \geq Z_{1-\alpha/2}$, z being the observed value of the test statistic. Alternatively, H_0 is rejected when the p -value is $p \leq \alpha$, where $p = 2[1 - \Phi(|z|)]$ and $\Phi(z)$ is the standard normal cumulative distribution function. To test H_0 against the one-sided lower or left-tailed alternative hypothesis $H_{1<}: \pi_1 < \pi_2$, one rejects H_0 when $z < Z_\alpha$ or the one-sided p -value is $p \leq \alpha$, where $p = \Phi(z)$. Likewise, to test H_0 against the one-sided upper or right-tailed alternative hypothesis $H_{1>}: \pi_1 > \pi_2$, one rejects H_0 when $z > Z_{1-\alpha}$ or the one-sided p -value is $p \leq \alpha$, where $p = 1 - \Phi(z)$.

Tests of this form based on an efficient estimate are asymptotically most powerful, or fully efficient against H_1 . It is important to note that such tests are constructed using the estimated standard error $\widehat{\sigma}_T(\theta_0)$ under the null hypothesis and not using an estimate under the alternative hypothesis, such as $\widehat{\sigma}_T(\theta_1)$ or $\widehat{\sigma}_T(\widehat{\theta})$. Asymptotically, a Z -test using these variance estimates obtained under the alternative also converges to $\mathcal{N}(0, 1)$ because each of these variance estimates also converges to $\sigma_T^2(\theta_0)$ under the null hypothesis H_0 . However, with small sample sizes, the size of the test may be inflated (or deflated), depending on whether the null hypothesis variance is under (or over)-estimated by these alternative variance estimates. Thus, in general, for a fixed sample size, one would expect a test based on the null hypothesis variance to have a true size closer to the desired significance level α than one based on a variance estimate under the alternative hypothesis, although both are asymptotically $\mathcal{N}(0, 1)$ under H_0 .

Thus, to test $H_0: \pi_1 = \pi_2$ in a 2×2 table, the asymptotic distribution under the null hypothesis of the risk difference presented in (2.26) and (2.27) leads to the usual expression for the large sample Z -test for two proportions based on the standardized deviate

$$Z = \frac{p_1 - p_2}{\hat{\sigma}_0} = \frac{p_1 - p_2}{\sqrt{p(1-p)[N/n_1 n_2]}}. \quad (2.83)$$

Since p_1 , p_2 , and p are asymptotically normally distributed, and since each has expectation π under H_0 , from Slutsky's theorem (A.45), $\hat{\sigma}_0 \xrightarrow{p} \sigma_0$ and asymptotically, $Z \xrightarrow{d} \mathcal{N}(0, 1)$ under H_0 . This Z -test is asymptotically fully efficient because it is based on the large sample estimate of the variance of \widehat{RD} under the null hypothesis H_0 . In Problem 2.8 we show that Z^2 in (2.83) equals the usual Pearson contingency chi-square statistic for a 2×2 table presented in the next section.

Another common approach to conducting a two-sided test of significance is to evaluate the $(1 - \alpha)$ -level confidence limits, computed as $T \pm Z_{1-\alpha/2} \hat{\sigma}_T(\hat{\theta})$, where $T = \hat{\theta}$ is consistent for θ . If these limits include θ_0 , then the test fails to reject H_0 at level α ; otherwise, one rejects H_0 at Type I error probability level α . This approach is equivalent to a two-sided Z -test using the *S.E.* of T , $\hat{\sigma}_T(\hat{\theta})$, estimated under the alternative hypothesis in the denominator of (2.82) rather than the *S.E.* estimated under the null hypothesis, $\hat{\sigma}_T(\theta_0)$. Since under the null hypothesis, $\hat{\sigma}_T^2(\hat{\theta}) \xrightarrow{p} \sigma_T^2(\theta_0)$, a test based on confidence intervals is asymptotically valid. However, the test based on the *S.E.* under the null hypothesis, $\hat{\sigma}_T(\theta_0)$ is, in general, preferred because the Type I error probability more closely approximates the desired level α . Thus, in cases where the variance of the test statistic depends on the expected value, as is the case for the test of proportions, there may be a discrepancy between the results of a significance test based on the Z -test of (2.83) and the corresponding two-sided confidence limits, in which case the test should be based on the former, not the latter. For the test of two proportions, since $\hat{\sigma}_T(\theta_0) > \hat{\sigma}_T(\hat{\theta})$ (see Lachin, 1981), it is possible that the $(1 - \alpha)$ -level confidence limits for the risk difference, or $\log(RR)$ or $\log(OR)$, would fail to include zero (implying significance) while the two-sided Z -test is not significant. In this case, one should use the Z -test and would fail to reject H_0 .

2.6.2 Unconditional Test

The Z -test in (2.83) is one of many common representations for the test of association in a 2×2 table, all of which are algebraically equivalent. Perhaps the most common is the usual Pearson contingency X_P^2 test for an $R \times C$ table with R rows and C columns. For a 2×2 table the test is

$$X_P^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}, \quad (2.84)$$

where n_{ij} is the observed frequency in the i th row and j th column as in (2.24) and \hat{E}_{ij} is the estimated expected frequency under the null hypothesis. For the case

of two independent groups, under $H_0: \pi_1 = \pi_2 = \pi$, an estimate of the common probability is provided by $\hat{\pi} = n_{1\bullet}/N$. This yields a table of expected frequencies within the cells of the table as

		Group		$n_{1\bullet}$
		1	2	
Response	+	$\hat{E}_{11} = n_{\bullet 1}\hat{\pi}$	$\hat{E}_{12} = n_{\bullet 2}\hat{\pi}$	
	-	$\hat{E}_{21} = n_{\bullet 1}(1 - \hat{\pi})$	$\hat{E}_{22} = n_{\bullet 2}(1 - \hat{\pi})$	$n_{2\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 2}$	N

These expected frequencies can also be derived under the hypothesis of independence between group and response. Let π_{ij} denote the probability of falling in the ij th cell, with row marginal probability $\pi_{i\bullet}$ and column marginal probability $\pi_{\bullet j}$. Then the hypothesis of independence specifies that $H_0: \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$. Since $\hat{\pi}_{i\bullet} = n_{i\bullet}/N$ and $\hat{\pi}_{\bullet j} = n_{\bullet j}/N$, then equivalently, $\hat{E}_{ij} = n_{i\bullet}n_{\bullet j}/N$.

Asymptotically, $X_P^2 \xrightarrow{d} \chi_1^2$, which designates the central chi-square distribution on 1 degree of freedom (df). There is only 1 degree of freedom because when any one of the cell frequencies is determined, the remainder are all obtained by subtraction.

For the 2×2 table, since the margins are fixed, then $|n_{ij} - \hat{E}_{ij}|$ is a constant for all cells of the table. Thus, the expression for X_P^2 reduces to

$$X_P^2 = \frac{(ad - bc)^2 N}{n_1 n_2 m_1 m_2}. \quad (2.85)$$

The test is directed to detect the two-sided alternative hypothesis $H_1: \pi_1 \neq \pi_2$. Thus, H_0 is rejected in favor of H_1 whenever $X_P^2 \geq \chi_{1-\alpha}^2$, the upper $1 - \alpha$ percentile of the central χ^2 distribution. As stated previously, it is readily shown that $Z^2 = X_P^2$. Since $\chi_{1-\alpha}^2 = (Z_{1-\alpha/2})^2$, it follows that the two-sided Z -test of H_0 versus H_1 using (2.83) is equivalent to the contingency chi-squared test that is inherently two-sided.

2.6.3 Conditional Mantel-Haenszel Test

Alternatively, as originally suggested by Mantel and Haenszel (1959) and extended by Mantel (1963), the test criterion could be based on the conditional central hypergeometric likelihood (2.73) rather than the product binomial. This likelihood involves a single random variable, the frequency a in the index cell of the 2×2 table. Thus, the *Mantel-Haenszel test* for the 2×2 table is most conveniently expressed in terms of the deviation of the observed value of the index frequency a from its expectation as

$$X_c^2 = \frac{[a - E(a)]^2}{V_c(a)}. \quad (2.86)$$

Using the factorial moments, or by direct solution, the moments of the central hypergeometric distribution (not merely their estimates) are

$$E(a) = \frac{n_1 m_1}{N} \quad (2.87)$$

and

$$V_c(a) = \frac{n_1 n_2 m_1 m_2}{N^2 (N - 1)} \quad (2.88)$$

(cf. Cornfield, 1956), where $V_c(a)$ is termed the *conditional variance*. The corresponding Z -statistic is

$$Z_c = \frac{a - E(a)}{\sqrt{V_c(a)}}. \quad (2.89)$$

Since a is the sum of *i.i.d.* Bernoulli variables, then asymptotically under H_0 , $Z_c \xrightarrow{d} \mathcal{N}(0, 1)$ which can be used for a one- or two-sided test. Thus, $X_c^2 \xrightarrow{d} \chi^2$ on 1 df under H_0 .

2.6.4 Cochran's Test

The unconditional contingency X_P^2 and Z -tests may also be expressed in terms of $[a - E(a)]$. Under the null hypothesis H_0 : $\pi_1 = \pi_2 = \pi$, from the unconditional product binomial likelihood (2.53), $E(a) = n_1 \pi$ and $V(a) = n_1 \pi(1 - \pi)$, each of which may be consistently estimated from (2.29) as $\widehat{E}(a) = n_1 p = n_1 m_1 / N$ and $\widehat{V}(a) = n_1 p(1 - p) = n_1 m_1 m_2 / N^2$. Likewise, $V(b) = n_2 \pi(1 - \pi)$ and $\widehat{V}(b) = n_2 p(1 - p)$. Since $m_1 = (a + b)$ is not fixed, then

$$a - \widehat{E}(a) = a - \frac{m_1 n_1}{N} = \frac{n_2 a - n_1 b}{N}. \quad (2.90)$$

Thus, the *unconditional variance* is

$$V_u = V[a - \widehat{E}(a)] = \frac{n_2^2 V(a) + n_1^2 V(b)}{N^2} = \frac{n_1 n_2 \pi(1 - \pi)}{N}, \quad (2.91)$$

which can be consistently estimated as

$$\widehat{V}_u = \widehat{V}[a - \widehat{E}(a)] = \frac{n_1 n_2 p(1 - p)}{N} = \frac{n_1 n_2 m_1 m_2}{N^3}. \quad (2.92)$$

Therefore, the unconditional test for the 2×2 table, X_u^2 , can be expressed as

$$X_u^2 = \frac{[a - \widehat{E}(a)]^2}{\widehat{V}_u}, \quad (2.93)$$

where $X_u^2 = X_P^2$ in (2.85). Likewise, the unconditional Z -test is

$$Z_u = \frac{a - \widehat{E}(a)}{\sqrt{\widehat{V}_u}}, \quad (2.94)$$

which is easily shown to equal the usual Z -test in (2.83). Thus, under H_0 , Z_u is asymptotically distributed as standard normal and X_u^2 as χ^2 on 1 df .

It is also instructive to demonstrate this result as follows. Asymptotically, assume that $n_1/N \rightarrow \xi$, $n_2/N \rightarrow (1 - \xi)$, ξ being the sample fraction for group 1. Also, since $m_1/N \xrightarrow{P} \pi$ and $m_2/N \xrightarrow{P} (1 - \pi)$ under H_0 , then from Slutsky's convergence theorem (A.45), $\widehat{E}(a)/n_1 \xrightarrow{P} \pi$ and $\widehat{V}(a)/n_1 \xrightarrow{P} (1 - \xi)\pi(1 - \pi)$. Since a is the sum of *i.i.d.* Bernoulli variables, then a/n_1 is asymptotically normally distributed. From Slutsky's theorem (A.43) and (A.44), it follows that Z_u is asymptotically distributed as standard normal.

The unconditional test in this form is often called *Cochran's test*. Cochran (1954a) described generalizations of the common chi-square test for the 2×2 table to various settings, most notably to the analysis of stratified 2×2 tables that is described in Chapter 4. Although Cochran's developments were generalizations of the Z -test of the form (2.83), his results are often presented as in (2.93) to contrast his test with that of Mantel and Haenszel.

Since

$$\widehat{V}_u = \frac{N-1}{N} V_c(a) < V_c(a) \quad (2.95)$$

then the conditional Mantel-Haenszel test tends to be slightly more conservative than the unconditional test. Clearly, the difference vanishes asymptotically, and thus the test using either the unconditional or conditional variance is often referred to as the *Cochran-Mantel-Haenszel (CMH)* test.

Example 2.8 Exact Inference Data and Neuropathy Clinical Trial (continued)

For the small sample size data in Example 2.7, $a = 7$, $E(a) = (19 \times 15)/29 = 9.8276$, $V_c(a) = (19 \times 10 \times 15 \times 14)/(29^2 \times 28) = 1.6944$ and $\widehat{V}_u = (28/29)V_c(a) = 1.6360$. Thus, the contingency χ^2 test or Cochran's test yields $X_P^2 = X_u^2 = (2.8276)^2/1.6360 = 4.8871$, with a corresponding value $Z_u = -2.211$ and a one-sided left-tailed p -value of $p \leq 0.0135$. The Mantel-Haenszel test yields $X_c^2 = (2.8276)^2/1.6944 = 4.7186$, with Z_c of -2.172 and $p \leq 0.0149$ (one-sided). Both tests yield p -values less than that of the exact test ($p = 0.033$), slightly less so for the Mantel test.

For such small samples, with total $N = 29$ in this example, the large sample tests are not valid. They are computed only as a frame of reference to the exact test presented previously. In general, for a 2×2 table, the minimum expected frequency for all four cells of the table, the E_{ij} in (2.84) should be at least five (some suggest four) for the asymptotic distribution to apply with reasonable accuracy (Cochran, 1954a).

For the clinical trial data in Example 2.2, the conditions are clearly satisfied. Here $a = 53$, $E(a) = 46.5$, $V_c(a) = 12.5013$, and $\widehat{V}_u = 12.4388$. Thus, Cochran's test yields $X_u^2 = 3.3966$ with $p \leq 0.0653$, inherently two-sided. The Mantel-Haenszel test yields $X_c^2 = 3.3797$ and $p \leq 0.066$. The large sample Z -test in (2.83) yields a value $z = 0.13/0.0699 = 1.843$ that equals $\sqrt{3.3966}$ from Cochran's test. These tests are not significant at the $\alpha = 0.05$ significance level two-sided.

2.6.5 Likelihood Ratio Test

Another test that is computed by many software packages, such as SAS, is the likelihood ratio or G^2 -test. Like the Pearson test, this test arises from consideration of the null hypothesis of statistical independence of the row and column factors. Using the notation of Section 2.6.2, under H_1 , the likelihood is a multinomial with a unique probability for each cell π_{ij} that can be estimated as $\hat{\pi}_{ij} = p_{ij} = n_{ij}/N$. Under H_0 , the likelihood is again a multinomial with cell probabilities that satisfy H_0 : $\pi_{ij0} = \pi_{i\bullet}\pi_{\bullet j}$, where the joint probabilities are estimated from the sample marginal proportions to yield the estimated expected frequencies \hat{E}_{ij} as in (2.84). Using $G^2 = -2 \log[L(\hat{\pi}_{ij0})/L(\hat{\pi}_{ij})]$ for the 2×2 table yields

$$G^2 = 2 \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \log \left[\frac{n_{ij}}{\hat{E}_{ij}} \right]. \quad (2.96)$$

Since this is a likelihood ratio test (see Section A.7.2), then asymptotically $G^2 \xrightarrow{d} \chi^2$ on 1 df . For the 2×2 table, this expression reduces to

$$\chi^2 = 2 \log \left[\frac{a^a b^b c^c d^d N^N}{n_1^{n_1} n_2^{n_2} m_1^{m_1} m_2^{m_2}} \right]. \quad (2.97)$$

The principal application of this test is to the analysis of log-linear models for multidimensional contingency tables. Further, the G^2 -test has been shown to converge to the asymptotic chi-square distribution at a slower rate than the contingency chi-square (or Cochran's) test (cf. Agresti, 1990). The relative rates of convergence of Cochran's versus the Mantel-Haenszel test have not been explored. However, given that the exact test is the yardstick against which the other tests have been compared, and given that the Mantel-Haenszel test is based on the normal approximation to the central hypergeometric distribution, one might expect that the Mantel test should be preferred.

2.6.6 Test-Based Confidence Limits

In some cases we may wish to construct a confidence interval for a parameter θ based on a consistent estimator $\hat{\theta}$, but where no convenient estimator of the variance of the estimate $\hat{V}(\hat{\theta}) = \hat{\sigma}_{\hat{\theta}}^2$ exists. Extending the Wilson (1927) limits for a proportion described in Section 2.1.3.4, suppose that a simple 1 df chi-square test, say X^2 , is available to test $H_0: \theta = 0$ versus $H_1: \theta \neq 0$. If we could estimate $V(\hat{\theta})$, we could construct an estimate-based test of the form $Q^2 = \hat{\theta}^2 / \hat{\sigma}_{\hat{\theta}}^2$. Thus, a test-based estimate of $\sigma_{\hat{\theta}}^2$ can be obtained by equating $Q^2 = X^2$ and solving for $\hat{\sigma}_{\hat{\theta}}^2$ to obtain

$$\hat{\sigma}_{\hat{\theta}}^2 = \frac{\hat{\theta}^2}{X^2} \quad (2.98)$$

and $S.E.(\hat{\theta}) = \hat{\theta}/X$. From this, the usual large sample confidence limits are obtained as $\hat{\theta} \pm Z_{1-\alpha/2} \hat{\theta}/X$.

Generalizations of this approach are widely used. For example, Miettinen (1976) used this approach to derive confidence limits for the Mantel-Haenszel estimate of the common log odds ratio for a set of 2×2 tables, as described in Chapter 4. However, the X^2 test usually employs an estimate of the variance of the test statistic under the null hypothesis, in which case this approach provides an estimate of the variance of $\hat{\theta}$ under the null hypothesis as well. This leads to some inaccuracy in the coverage probabilities of the confidence limits. See Greenland (1984) for a review of the controversies regarding this approach. In general, when appropriate alternatives exist, this approach should not be the first choice. However, it can be employed to obtain approximate confidence limits for parameters in other cases where the variance of the statistic under the alternative may not be tractable.

Example 2.9 *Neuropathy Clinical Trial (continued)*

Consider the log odds ratio for a 2×2 table. Suppose that we wished to estimate the variance of the sample log odds ratio $\hat{\theta}$ using this approach. The Mantel-Haenszel test X_c^2 in (2.86) is an asymptotically fully efficient test for $H_0: \theta = 0$ versus $H_1: \theta \neq 0$ since it is based on the large sample approximation to the hypergeometric distribution for which the odds ratio is the noncentrality parameter. Thus, the $S.E.(\hat{\theta}) = \hat{\theta}/X_c$.

For the data in Example 2.2, the sample log odds ratio is $\hat{\theta} = 0.52561$, and the value of the Mantel-Haenszel test is $X_c^2 = 3.37966$. Thus, the $S.E. = 0.52561/\sqrt{3.37966} = 0.2859$, which is nearly identical to the large sample $S.E. = \hat{\sigma}_{\hat{\theta}} = 0.2860$ obtained from (2.47). Thus, in this case the resulting confidence limits are nearly identical to those presented previously.

2.6.7 Continuity Correction

The large sample tests X_P^2 , X_c^2 , and X_u^2 , presented in (2.85), (2.86), and (2.93), respectively, and their corresponding Z -tests are based on the normal approximations to the binomial and the central hypergeometric distributions in which these discrete distributions are approximated by the smooth, continuous normal distribution. For small sample sizes, the adequacy of the approximation is vastly improved through use of the continuity correction. For example, the central hypergeometric distribution for a 2×2 table with the fixed margins of the example in (2.79) could be plotted as a histogram with masses of probability at discrete values $5 \leq a \leq 15$. For the observed table with $a = 7$, the exact one-sided lower tailed p -value is obtained as the sum of the histograms for $5 \leq a \leq 7$. Clearly, to approximate this tail area under the smooth normal curve, we would use $P(a \leq 7.5)$ so as to capture the total probability associated with $a = 7$.

Thus, the continuity-corrected conditional and unconditional tests X_c^2 in (2.86) and X_u^2 in (2.93), and their corresponding Z -tests in (2.89) and (2.94), would employ

$$|a - E(a)| - 1/2 \quad (2.99)$$

in the numerator in place of simply $[a - E(a)]$.

There has been substantial debate over the utility of the continuity correction. Those interested are referred to Conover (1974) and the discussion thereto by Starmer

et al. (1974), Mantel (1974), and Miettinen (1974a), and the references therein. Mantel and Greenhouse (1968) and Mantel (1974) have argued that the exact test p -value from the conditional hypergeometric distribution should be used as the basis for the comparison of the continuity corrected and uncorrected tests. In their comparisons, the continuity corrected statistic is preferred. Conover (1974) points out that this advantage principally applies when either margin is balanced ($n_1 = n_2$ or $m_1 = m_2$), but not otherwise.

However, Grizzle (1967) and Conover (1974), among others, have shown that the uncorrected statistic more closely approximates the complete exact distribution, not only the tails. This is especially true if one considers the significance level obtained from the marginal expectation of the corrected and uncorrected test statistics on sampling from two binomial distributions without the second margin (the m_1 and m_2) fixed. Starmer et al. (1974) argued that since the exact significance level is some value $\alpha^* \leq \alpha$ due to the discreteness of the binomial and hypergeometric distributions, then the basis for comparison should be the uniformly most powerful test under both product-binomial and hypergeometric sampling. This is the randomized exact test of Tocher (1950), wherein randomization is employed to yield an exact Type I error probability exactly equal to α . Compared to the randomized exact test tail areas, the continuity corrected tests are far too conservative.

Tocher's test, however, is not widely accepted. Thus, the continuity correction is most useful with small sample sizes. As the sample size increases, and the central limit theorem takes hold, the uncorrected statistic is preferred, except perhaps in those cases where one of the margins of the table is perfectly balanced. Also, with the advent of modern computer systems, exact computations are readily performed with small sample sizes where the continuity correction is most valuable. In general, therefore, the continuity correction should be considered with small sample sizes for cases where exact computations are not available.

Example 2.10 *Exact Inference Data and Neuropathy Clinical Trial (continued)*

For the 2×2 table with small sample sizes in Example 2.7, the continuity-corrected conditional Z_c -test is $[-2.8276| - 0.5]/\sqrt{1.69444}$. Retaining the sign of the difference, this yields $Z_c = -2.3276/1.3017 = -1.788$ with a one-sided lower p -value of $p \leq 0.0369$, which compares more favorably to the one-sided exact p -value of $p \leq 0.0329$ (see Example 2.7). The uncorrected Z -test yields a p -value of $p \leq 0.0149$, much smaller than the exact p -value.

For the clinical trial data in Example 2.2 with much larger sample sizes, the two-sided p -values for the uncorrected and corrected conditional X_c^2 are 0.06601 and 0.0897. The corrected statistic p -value compares favorably with the exact two-sided p -value of 0.0886, principally because the group margin is balanced, $n_1 = n_2 = 100$.

Now consider the following slight modification of the data in Example 2.2, where the margins are unbalanced:

		Group		
		1	2	
Response	+	60	36	96
	-	53	54	107
		113	90	203

(2.100)

The pivotal quantities are virtually unchanged [$OR = 1.698$, $a - E(a) = 6.56$, and $V_c(a) = 12.5497$], so that the uncorrected and corrected X_c^2 p -values are 0.064 and 0.087, nearly the same as those for the example with the margins balanced. Nevertheless, the exact two-sided p -value changes to 0.0677, so that the corrected X_c^2 is too conservative.

In general, therefore, with small sample sizes, the exact calculation is preferred. With large sample sizes, the continuity correction should not be used except in cases where one of the two margins of the table is perfectly balanced, or nearly so.

2.6.8 Establishing Equivalence or Noninferiority

2.6.8.1 Equivalence All of the preceding large sample tests are used to establish the superiority of one group versus another, either in terms of a higher fraction of a positive or favorable outcome, or a lesser fraction of a negative or adverse outcome. Each test is evaluated under the null hypothesis $H_0: \pi_1 = \pi_2$ and is designed to provide a basis for rejecting H_0 in favor of H_1 that the probabilities differ. In some cases, however, the objective of a study may be to demonstrate that two groups (treatments, exposures, etc.) are equivalent within some margin of error. Such equivalence is usually established through the construction of an appropriate confidence interval.

For the case of two proportions the equivalence margin can be specified in terms of a risk difference, relative risk, or odds ratio. To begin, consider the risk difference of a positive outcome for a new therapy with probability π_1 versus an existing standard of care with probability π_2 . The margin would then specify the limit on the difference $\Delta_E = |\pi_1 - \pi_2|$ that would constitute equivalence. Equivalence would be established if the $(1 - \alpha/2)$ -level confidence limits on this difference are contained with the margins of equivalence, or

$$(\hat{\Delta}_L, \hat{\Delta}_U) = (p_1 - p_2) \pm Z_{1-\alpha/2} \hat{\sigma}_1 \in (-\Delta_E, \Delta_E), \quad (2.101)$$

where $\hat{\sigma}_1^2$ is the estimated variance under the alternative in (2.28).

Note that unlike a test of significance, these computations employ the variance of the statistic evaluated under the alternative rather than the null hypothesis. Also note that the above specifies a symmetric margin of error. However, the developments herein are readily generalized to allow for asymmetric margins of error $(\Delta_{EL}, \Delta_{EU})$ where $\Delta_{EL} \neq -\Delta_{EU}$.

The criterion for equivalence is inherently two-sided. In fact, the criterion in (2.101) for the risk difference can be expressed in terms of two simultaneous one-sided tests of the hypotheses $H_0: \Delta = -\Delta_E$ and $H_0: \Delta = \Delta_E$, each versus $H_1: \Delta = 0$, with the respective rejection criteria

$$\begin{aligned} Z &= \frac{(p_1 - p_2) + \Delta_E}{\hat{\sigma}_1} \geq Z_{1-\alpha/2} \\ Z &= \frac{(p_1 - p_2) - \Delta_E}{\hat{\sigma}_1} \leq Z_{\alpha/2} \end{aligned} \quad (2.102)$$

(Schuirmann, 1987). Equivalence is established when both rejection criteria are met, which occurs only when the observed difference is in the neighborhood of zero. These equations in turn satisfy the criterion

$$\frac{(p_1 - p_2) \pm \Delta_E}{\hat{\sigma}_1} \in (Z_{\alpha/2}, Z_{1-\alpha/2}), \quad (2.103)$$

that is equivalent to (2.101).

Likewise, the criterion for equivalence could be specified in terms of asymmetric confidence limits on a relative risk obtained using the log transformation as in (2.42). In this case the margin is specified on the log relative risk, such as $\Delta_E = \theta_E = \log(\pi_1/\pi_2)$ ($\theta_E > 0$), and equivalence is satisfied if

$$(\widehat{RR}_L, \widehat{RR}_U) = \exp(\widehat{\theta}_L, \widehat{\theta}_U) \in (e^{-\theta_E}, e^{\theta_E}). \quad (2.104)$$

When described using two one-sided tests as in (2.102), the criterion for the $\log(RR)$ can be expressed as

$$\frac{\log(\widehat{RR}) \pm \theta_E}{\hat{\sigma}_1} \in (Z_{\alpha/2}, Z_{1-\alpha/2}). \quad (2.105)$$

where $\hat{\sigma}_1^2 = \widehat{V}[\log(\widehat{RR})]$ is as expressed in (2.38). This is equivalent to (2.104).

Similarly, the equivalence criterion could be specified in terms of a log odds ratio.

2.6.8.2 Noninferiority In most cases, however, the two-sided margin is unnecessary or irrelevant. Rather, the question may be whether the new therapy is not inferior to the standard, termed the *evaluation of noninferiority*. For example, if group 1 receives a new therapy and group 2 the standard of care, then noninferiority would mean that the new therapy is not worse in promoting a positive outcome than the standard within a margin of error. For the risk difference, stated in terms of a positive outcome where a positive difference is beneficial, and with a specified margin of error Δ_E , noninferiority is established if

$$\widehat{\Delta}_L = (p_1 - p_2) - Z_{1-\alpha} \hat{\sigma}_1 > -\Delta_E. \quad (2.106)$$

Since the criterion is inherently one-sided, then the one-sided confidence limit is often employed.

Table 2.2 SAS Program for analysis of 2×2 tables.

```

title1 'SAS PROC FREQ Analysis of 2x2 tables';
data one; input a b c d; cards;
53 40 47 60
;
title2 'Neuropathy Clinical Trial, Example 2.2';
data two; set one;
group=1; response=1; x=a; output;
group=2; response=1; x=b; output;
group=1; response=2; x=c; output;
group=2; response=2; x=d; output;
proc freq; tables group*response / nopercent nocol all riskdiff;
exact or; weight x;
run;

```

If the noninferiority criterion is specified in terms of a log relative risk (or log odds ratio) θ , then noninferiority is established if

$$\widehat{RR}_L = \exp(\widehat{\theta}_L) > e^{-\Delta_E} \quad (2.107)$$

using the one-sided lower confidence limit on θ . In other cases, the noninferiority margin may be specified in terms of the risk of an adverse outcome, in which case a lower proportion is favorable. Then the upper confidence limit would be employed, such as

$$\begin{aligned} \widehat{\Delta}_U &= (p_1 - p_2) + Z_{1-\alpha} \widehat{\sigma}_1 < \Delta_E, \quad \text{or} \\ \widehat{RR}_U &= \exp(\widehat{\theta}_U) < e^{\Delta_E}. \end{aligned} \quad (2.108)$$

Even though the hypothesis of noninferiority is specified one-sided, in some cases, such as is often required by the FDA, the two-sided confidence limit is employed, or a one-sided limit at level $1 - \alpha/2$, in which case the above computations would employ $Z_{1-\alpha/2}$.

Example 2.11 Neuropathy Clinical Trial (continued)

Again consider the neuropathy clinical trial data where group 1 receives the new therapy, but now assuming that group 2 receives the standard therapy rather than a placebo. Assume that an equivalence criterion is prespecified as a risk difference within ± 0.10 or $\Delta_E = 0.10$. As shown in Example 2.3, the 95% confidence limits on the risk difference are $(-0.0071, 0.2671)$. Since these are not contained within the interval $(-0.1, 0.1)$, equivalence is not established.

Using the same data, assume that a noninferiority criterion was prespecified if the relative risk was no less than 0.9, or a margin of 10% using a two-sided calculation, as may be required by the FDA. As shown in Example 2.4, the lower two-sided 95% confidence limit on the relative risk is 0.9788, that is greater than 0.9 in which

Table 2.3 2×2 table for neuropathy clinical trial.

SAS PROC FREQ Analysis of 2x2 tables
Neuropathy Clinical Trial, Example 2.2

TABLE OF GROUP BY RESPONSE

GROUP		RESPONSE		Total
Frequency	Row Pct	1	2	
1	53	47		100
	53.00	47.00		
2	40	60		100
	40.00	60.00		
Total		93	107	200

case noninferiority could be claimed. The calculation could also be done in terms of $\widehat{\theta}_L = \log(\widehat{RR})$ using $\Delta_E = \log(1.1)$.

This example illustrates why an assessment of equivalence is rarely used. The data show that the new therapy is clearly not inferior to the standard. In fact, the new therapy might have been declared superior had a one-sided test been prespecified since the one-sided p -value from Example 2.8 is 0.034. This trend toward superiority then leads to failure to establish equivalence.

2.7 SAS PROC FREQ

Many of the computations described in this section are provided by the Statistical Analysis System (SAS) procedure for the analysis of cross tabulations of frequency data, PROC FREQ. These include the computation of confidence limits for a single proportion using the *binomial* option.

To conduct an analysis of the data from a 2×2 table such as those presented herein, the simple SAS program presented in Table 2.2 could be employed. Each 2×2 table is input in terms of the frequencies in each cell, as in (2.24). However, in order to analyze the 2×2 table in SAS, it is necessary to create a data set with four observations, one per cell of the 2×2 table, and three variables: one representing group and another response, each with distinct numerical or categorical values, and one representing the frequency within that cell. Then PROC FREQ is called using this data set. Tables 2.3 to 2.6 present the results of the analysis.

The 2×2 table is presented in Table 2.3. The SAS program uses the *table* statement for *group*response*, so that *group* forms the rows and *response* the columns. This is necessary because SAS computes the relative risks across rows rather than across columns as presented throughout this book. The *nocol* and *nopercent* options suppress printing of irrelevant percentages.

The *all* option requests statistical tests, the measures of the strength of the association between *group* and *response*, the odds ratio and relative risks. These results are presented in Table 2.4. The various test statistics are presented, followed by various measures of the degree of association between *group* and *response*, most having been deleted from the output since they are not discussed herein. Among these, the uncertainty coefficient for *response* (columns) given *group* (rows), designated as $U(C|R)$, is a direct R^2 measure like the fraction variation explained in a linear regression model. It is equivalent to the entropy R^2 from a logistic regression model that is described in Chapter 7.

The *riskdiff* option generates the "column 1 risk estimates," that are the conditional proportions of a positive response within each group (row). The output also contains the risk estimates (proportions) within column 2 (a negative response), that are not of interest and thus are not shown. In addition to the large sample confidence limits for each probability, the exact limits for each are also presented. The program then computes the difference in the proportions between the two rows (groups), its large sample (asymptotic) standard error (ASE), and confidence limits.

SAS then presents estimates of the relative risk for row 1/row 2 (group 1 vs. group 2 in this instance). The odds ratio is labeled as *case-control (odds ratio)* because, as will be shown in Chapter 5, the relative risk in a case-control study may be approximated by the odds ratio. This is followed by the *cohort (col1 risk)* that is the relative risk of the proportions in the first column. The other is the *cohort (col2 risk)* that is the ratio of the response proportions in the second column. Of the two, for this example the *col1* relative risk is the preferred statistic because the first column represents the positive category of interest. For each measure, the asymmetric confidence limits based on the δ -method are presented, using the log of the odds ratio and the log of the relative risks. The *exact or* option also generates the exact confidence limits for the odds ratio.

The *all* or *CMH* options also generate the Cochran-Mantel-Haenszel (CMH) statistics that are presented in Table 2.5. For a single 2×2 table such as this, the results are equivalent to the Mantel-Haenszel test presented in Table 2.4. The distinction among the different CMH tests is described in Section 2.10 for polychotomous data. Then the measures of relative risk for a case-control study (the odds ratio) and the *col1* and *col2* relative risks are presented. In each case the *Mantel-Haenszel* and *logit* estimates are presented with 95% confidence limits. For a single 2×2 table, the Mantel-Haenszel and logit point estimates are the same as the simple estimates described herein. For the *Mantel-Haenszel* estimates, the confidence limits are the test-based confidence limits described in (2.98). For the *logit* estimates, the confidence limits are the asymmetric limits based on the δ -method estimated variance. Note that SAS refers to these as logit estimates, although the logit only applies to the odds ratio.

Table 2.4 2×2 table statistics.

STATISTICS FOR TABLE OF GROUP BY RESPONSE

Statistic	DF	Value	Prob
Chi-Square	1	3.397	0.065
Likelihood Ratio Chi-Square	1	3.407	0.065
Continuity Adj. Chi-Square	1	2.894	0.089
Mantel-Haenszel Chi-Square	1	3.380	0.066
Fisher's Exact Test (Left)			0.977
		(Right)	0.044
		(2-Tail)	0.089

Statistic	Value	ASE
Uncertainty Coefficient C R	0.012	0.013

Column 1 Risk Estimates

(Asymptotic) 95%

	Risk	ASE	(Asymptotic) 95% Confidence Limits	
Row 1	0.530	0.050	0.432	0.628
Row 2	0.400	0.049	0.304	0.496
Difference	0.130	0.070	-0.007	0.267
(Row 1 - Row 2)				

(Exact) 95%
Confidence Limits

Row 1	0.428	0.631
Row 2	0.303	0.503

Case-Control (Odds ratio)	1.691	0.966	2.963
Cohort (Col1 Risk)	1.325	0.979	1.794
Cohort (Col2 Risk)	0.783	0.602	1.019

Odds Ratio (Case-Control Study)

Exact Conf. Limits

95% Lower Conf Limit 0.9300
95% Upper Conf Limit 3.0805

Table 2.5 Cochran-Mantel-Haenszel statistics.

SAS PROC FREQ Analysis of 2x2 tables
Neuropathy Clinical Trial, Example 2.2

SUMMARY STATISTICS FOR GROUP BY RESPONSE

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	3.380	0.066
2	Row Mean Scores Differ	1	3.380	0.066
3	General Association	1	3.380	0.066

Estimates of the Common Relative Risk (Row1/Row2)

95%

Type of Study	Method	Value	Confidence Bounds	
Case-Control (Odds Ratio)	Mantel-Haenszel	1.691	0.966	2.962
	Logit	1.691	0.966	2.963
Cohort (Col1 Risk)	Mantel-Haenszel	1.325	0.982	1.789
	Logit	1.325	0.979	1.794
Cohort (Col2 Risk)	Mantel-Haenszel	0.783	0.604	1.016
	Logit	0.783	0.602	1.019

The confidence bounds for the M-H estimates are test-based.

Exact limits for the odds ratio, risk difference, and relative risk (risk ratio) are provided by StatXact.

Finally, the *riskdiff* option also provides options for the assessment of the hypotheses of equivalence or noninferiority using a specification in terms of a risk difference with a specified margin. However, the default specification of $\alpha = 0.05$ results in 90% equivalence limits rather than the usual 95% limits. Thus, an additional call of the procedure is required, such as the following:

```
proc freq; tables group*response / nopercent nocol
  alpha = 0.025 riskdiff (equivalence margin = 0.1);
```

This will provide the desired 95% limits for the test of equivalence but 97.5% limits for all other quantities. When applied to the neuropathy clinical trial data as in

Table 2.6 Assessment of equivalence for the neuropathy clinical trial data.

Statistics for Table of group by response			
Equivalence Analysis for the Proportion (Risk) Difference			
H0: $P_1 - P_2 \leq$ Lower Margin or \geq Upper Margin			
Ha: Lower Margin $< P_1 - P_2 <$ Upper Margin			
Lower Margin = -0.1 Upper Margin = 0.1 Wald Method			
Proportion Difference		ASE (Sample)	
0.1300		0.0699	
Two One-Sided Tests (TOST)			
Test	Z	P-Value	
Lower Margin	3.2887	Pr > Z	0.0005
Upper Margin	0.4290	Pr < Z	0.6660
Overall			0.6660
Equivalence Limits		95% Confidence Limits	
-0.1000	0.1000	-0.0071	0.2671

Example 2.11, the results in Table 2.6 are obtained that include the two one-sided test values as in (2.102) and the 95% confidence limits on the difference.

The procedure also provides an assessment of noninferiority for the risk difference. However, it does not provide either an assessment of equivalence or noninferiority for other measures, such as the relative risk. However, this could be done directly from the confidence limits on the parameter of interest.

2.8 OTHER MEASURES OF DIFFERENTIAL RISK

There are many other possible measures of the difference in risk between two groups of subjects. Three in particular are becoming more widely used in the presentation of research results: the attributable risk, the population attributable risk, and the number needed to treat. The large sample distribution of each, and a large sample estimate of the variance of each, is readily obtained from the results presented in this chapter. Details are left to the problems.

2.8.1 Attributable Risk Fraction

Consider an observational study comparing the risk of an adverse outcome, say, the incidence of a disease, in a sample of individuals exposed to a putative risk factor (E or group 1) versus that in a sample of nonexposed individuals (\bar{E} or group 2). In this setting the risk difference $\pi_1 - \pi_2$ is at times termed the *attributable risk* since it is a measure of the absolute excess risk of the outcome that can be attributed to the exposure in the population. However, a more useful measure is the attributable risk

(*AR*) fraction

$$AR = \frac{\pi_1 - \pi_2}{\pi_2} = RR - 1 \quad (2.109)$$

(MacMahon and Pugh, 1970), where $RR = \pi_1/\pi_2$ is the relative risk of the disease among exposed versus nonexposed individuals, and where it is assumed that exposure leads to an increase in the risk of the disease, $\pi_1 > \pi_2$. This is a measure of the fractional or proportionate increase in the risk of the disease caused by exposure to the risk factor. Since this is a simple function of the RR , asymmetric confidence limits on the AR are readily obtained from the asymmetric limits on RR computed using the log transformation as described in Section 2.3.2.

2.8.2 Population Attributable Risk

The attributable risk fraction is the proportionate excess risk associated with the exposure. However, it does not account for the prevalence of the exposure in the population, and thus is lacking from the perspective of the overall effects on the public health. Thus, Levin (1953) proposed the *population attributable risk fraction (PAR)*, which is the proportion of all cases of the disease in the population that are attributable to exposure to the risk factor. Equivalently, it is the proportion of all cases of the disease in the population that would be avoided if the adverse exposure could be eliminated in the population. For example, it addresses a question such as "What fraction of lung cancer deaths could be eliminated if smoking were completely eliminated in the population?" Such questions can be answered when we have a random sample of N individuals from the general population such that the fraction exposed in the sample is expected to reflect the fraction exposed in the population. This quantity was also termed the *etiologic fraction* by Miettinen (1974b), and was studied by Walter (1975, 1976).

The *PAR* can be derived as follows: In the general population, let $\alpha_1 = P(E)$ be the fraction exposed to the risk factor, and let the complement $\alpha_2 = 1 - \alpha_1 = P(\bar{E})$ be the proportion not exposed. Among those exposed, the fraction who develop the disease, say D , is $\pi_1 = P(D | E)$; and among those not exposed, the proportion who do so is $\pi_2 = P(D | \bar{E})$. From these quantities we can then describe the probability of cases of the disease among those exposed (E) versus not (\bar{E}) as in the following table:

		Group		π_*
		E	\bar{E}	
D	$\alpha_1\pi_1$	$\alpha_2\pi_2$	π_*	$1 - \pi_*$
	$\alpha_1(1 - \pi_1)$	$\alpha_2(1 - \pi_2)$	$1 - \pi_*$	
	α_1	α_2	1.0	

The total probability of the disease in the population is $\pi_* = \alpha_1\pi_1 + \alpha_2\pi_2$. Of these, $\alpha_1\pi_1$ is the fraction associated with exposure to the risk factor and $\alpha_2\pi_2$ that associated with nonexposure. However, if exposure to the risk factor could be eliminated in the population, then the fraction previously exposed (α_1) would have the same probability of developing the disease as those not exposed (π_2), so that the

fraction with the disease would be $\alpha_1\pi_2$. Thus, the population attributable risk is the fraction of all cases of the disease that would be prevented if the exposure to the risk factor were eliminated in the population:

$$\begin{aligned} PAR &= \frac{\alpha_1\pi_1 - \alpha_1\pi_2}{\alpha_1\pi_1 + \alpha_2\pi_2} = \frac{\alpha_1(\pi_1 - \pi_2)}{\pi_*} = \frac{\pi_* - \pi_2}{\pi_*} \\ &= \frac{\alpha_1(RR - 1)}{\alpha_1RR + \alpha_2} = \frac{\alpha_1(RR - 1)}{1 + \alpha_1(RR - 1)}. \end{aligned} \quad (2.110)$$

Under H_0 , $RR = 1$ and $PAR = 0$.

For example, consider that we have a random sample of N individuals from the general population, of whom n_1 have been exposed to the risk factor. Of these, $n_1\pi_1$ are expected to develop the disease. However, if these n_1 individuals were not exposed to the risk factor, then only $n_1\pi_2$ of them would be expected to develop the disease. Thus, the expected number of disease cases saved or eliminated if the exposure could be eliminated is $n_1\pi_1 - n_1\pi_2$. Since π_*N is the total expected number of cases, then given n_1 of N exposed individuals, the proportion of cases eliminated or saved by complete eradication of the exposure is

$$PAR = \frac{n_1\pi_1 - n_1\pi_2}{\pi_*N}. \quad (2.111)$$

In such a random sample, $\hat{\alpha}_1 = n_1/N$, $\hat{\pi}_1 = p_1$, $\hat{\pi}_2 = p_2$, and $\hat{\pi}_* = m_1/N$. Substituting into (2.110), the PAR can be consistently estimated as

$$\widehat{PAR} = \frac{\hat{\alpha}_1(\widehat{RR} - 1)}{1 + \hat{\alpha}_1(\widehat{RR} - 1)} = \frac{a - n_1p_2}{m_1}. \quad (2.112)$$

Implicit in this expression is the requirement that the sample fraction of exposed individuals provides an unbiased estimate of the fraction exposed in the population; that is, $\hat{\alpha}_1 = \widehat{P}(E) = n_1/N$ is unbiased for $\alpha_1 = P(E)$.

Walter (1976) described the computation of confidence limits for the PAR based on a large sample estimate of the variance of the \widehat{PAR} . However, because the PAR is a probability bounded by $(0,1)$ when $\pi_1 \geq \pi_2$, it is preferable that confidence limits be based on the asymptotic distribution of the logit of \widehat{PAR} . If we assume that α_1 is known (or fixed), using the δ -method (see Problem 2.10) it is readily shown that the logit of \widehat{PAR} is asymptotically normally distributed with expectation $\text{logit}(PAR)$ and with a large sample variance

$$V \left[\log \frac{\widehat{PAR}}{1 - \widehat{PAR}} \right] = \frac{\pi_1}{(\pi_1 - \pi_2)^2} \left[\frac{n_2\pi_2(1 - \pi_1) + n_1\pi_1(1 - \pi_2)}{n_1n_2\pi_2} \right]. \quad (2.113)$$

This quantity does not involve the population exposure probability α_1 . Using Slutsky's theorem the variance can be consistently estimated as

$$\widehat{V} \left[\log \frac{\widehat{PAR}}{1 - \widehat{PAR}} \right] = \frac{p_1}{(p_1 - p_2)^2} \left[\frac{n_2p_2(1 - p_1) + n_1p_1(1 - p_2)}{n_1n_2p_2} \right]. \quad (2.114)$$

From this expression, the large sample $(1 - \alpha)$ -level confidence limits for the logit and the asymmetric confidence limits on \widehat{PAR} are readily obtained. Leung and Kupper (1981) derived a similar estimate of $V[\text{logit}(\widehat{PAR})]$ based on the estimate in (2.112), which also involves the sample estimate of α_1 .

While the PAR is frequently used in the presentation of epidemiologic research, it may not be applicable in many instances. As cited by Walter (1976), among others, the interpretation of the PAR as a fraction of disease cases attributable to exposure to a risk factor requires the assumption that the risk factor is, in fact, a cause of the disease and that its removal alters neither the distribution of other risk factors in the population, nor their effects on the prevalence of the disease. Rarely can it be asserted that these conditions apply exactly.

Example 2.12 Coronary Heart Disease in the Framingham Study

Walter (1978) described the attributable risk using data from Cornfield (1962). In an early analysis of the Framingham study, Cornfield (1962) showed that the risk of coronary heart disease (CHD) during six years of observation among those with and without an initial serum cholesterol level ≥ 220 mg/dL was

		Serum Cholesterol		92	1237	1329	(2.115)
		≥ 220	< 220				
Response	CHD	72	20				
	CHD-free	684	553				
		756	573				

(reproduced with permission). Among those with elevated cholesterol (≥ 220 mg/dL), 9.52% of subjects (p_1) developed CHD versus 3.49% (p_2) among those without an elevated cholesterol (< 220 mg/dL), with a relative risk of 2.729 and 95% C.I. (1.683, 4.424), that is highly significant at $p < 0.001$. Thus, the attributable risk in (2.109) is 1.729 with a 95% C.I. of (0.683, 3.424), which indicates that the incidence of CHD is 173% greater among those with elevated cholesterol values (95% C.I. 68, 342%).

The estimated population attributable risk is $\widehat{PAR} = 0.4958$, which indicates that nearly 50% of the CHD events in the population may be attributable to elevated serum cholesterol levels. Using the equations provided by Walter (1978) based on an estimate of the S.E. of the \widehat{PAR} , the 95% confidence limits are (0.3047, 0.6869). The logit of the \widehat{PAR} is -0.01685 with an estimated variance from (2.114) of 0.15155. This yields 95% confidence limits on $\text{logit}(\widehat{PAR}) = (-0.7798, 0.7462)$. Taking the inverse logits or applying the logistic function yields asymmetric confidence limits on the PAR of (0.3144, 0.6783). In this example, since the \widehat{PAR} is close to 0.5, the two sets of limits agree closely; however, with values approaching 0 or 1, the logit limits are expected to be more accurate and remain bounded by (0,1).

2.8.3 Number Needed to Treat

The attributable risk and the population attributable risk are principally intended to aid in the interpretation of observational epidemiologic studies. For a randomized clinical trial of a new therapy, the number needed to treat (NNT) has been suggested as a summary of the effectiveness of the therapy with the aim of preventing an adverse outcome or promoting a favorable outcome (Laupacis et al., 1988; Cook and Sackett, 1995). The question is how many patients must be treated with the new therapy in order to prevent a single adverse outcome, or to promote a single favorable outcome. First, consider the case where the probability of a positive favorable outcome with the experimental treatment π_1 is greater than that with the control treatment, π_2 . Then the NNT satisfies the expression $NNT(\pi_1 - \pi_2) = 1$, so that

$$NNT = \frac{1}{\pi_1 - \pi_2} = \frac{1}{RD} \quad (2.116)$$

is the number needed to treat to yield a positive outcome in a single patient. Conversely, in the case where the probability of a negative outcome with the experimental treatment π_1 is less than that with the control treatment, π_2 , the NNT satisfies $NNT(\pi_2 - \pi_1) = 1$, so that $NNT = -1/RD$. Thus, confidence limits on NNT are readily obtained from confidence limits on the risk difference RD . In other settings, such as survival analysis, this concept may still be applied using the difference in cumulative incidence probabilities as of a fixed point in time.

Example 2.13 Cholesterol and CHD

Assume now that we could treat elevated cholesterol values effectively with a drug so that among the drug-treated patients the six-year incidence of CHD is $\pi_1 = 0.040$ whereas that among the controls, whose cholesterol remains elevated, is $\pi_2 = 0.095$. Then the number needed to treat with the new drug in order to prevent a single case of CHD is $1/0.055 = 18.18$ or 19 patients.

2.9 POLYCHOTOMOUS AND ORDINAL DATA

2.9.1 Multinomial Distribution and Large Sample Approximation

The preceding sections describe basic methods for the analysis of a binary categorical variable, where interest is focused on one of the two categories. Often the data of interest consist of a categorical variable with multiple ($C > 2$) categories, such as grades of severity of an illness. For a sample of N observations, the data consists of the frequencies within the C categories X_1, \dots, X_C , each with a corresponding probability π_1, \dots, π_C . The underlying distribution is multinomial and the probability of a given set of frequencies is provided by

$$P(x_1, \dots, x_C) = \frac{N!}{\prod_{j=1}^C x_j!} \prod_{j=1}^C \pi_j^{x_j}. \quad (2.117)$$

As shown in Example A.2, from the multivariate central limit theorem it follows that the vector of frequencies $\mathbf{x} = (x_1 \cdots x_C)^T$ is distributed as multivariate normal with expectation $N\boldsymbol{\pi} = N(\pi_1 \cdots \pi_C)^T$ and covariance matrix $N\boldsymbol{\Sigma}(\boldsymbol{\pi})$ with $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ as shown in (A.26) with elements

$$\begin{aligned}\sigma_{jj}^2 &= \pi_j(1 - \pi_j), \quad j = 1, \dots, C \\ \sigma_{jk} &= -\pi_j\pi_k, \quad j \neq k.\end{aligned}\quad (2.118)$$

Using $\mathbf{D}_\pi = \text{diag}(\pi_1 \cdots \pi_C)$ to denote the diagonal matrix in the event probabilities, then $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ is also expressed as

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}', \quad (2.119)$$

where $\boldsymbol{\pi}\boldsymbol{\pi}'$ is the outer product of the vector $\boldsymbol{\pi}$. It also follows that $\mathbf{p} = (p_1 \cdots p_C)^T$ is distributed as multivariate normal with expectation $\boldsymbol{\pi}$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\pi})/N$.

However, the distribution is degenerate, owing to the linear constraint $\sum_{j=1}^C x_j = N$, or $\sum_{j=1}^C p_j = 1$, so that $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ is singular of rank $C - 1$. Thus, to characterize the joint distribution of frequencies or proportions for a given variable, it is customary that the distribution of $C - 1$ of the categories be employed. Let $\tilde{\mathbf{p}} = (p_1 \cdots p_{C-1})^T$, $\tilde{\boldsymbol{\pi}} = (\pi_1 \cdots \pi_{C-1})^T$, and $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\pi})$ be the $(C - 1)(C - 1)$ submatrix of $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ obtained by removing the final row and column. Then

$$\tilde{\mathbf{p}} \stackrel{d}{\approx} \mathcal{N}(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\pi})/N) \quad (2.120)$$

of order and rank $C - 1$.

This distribution would be used to test the null hypothesis that the sample was drawn from a distribution with a set of prespecified category probabilities, or H_0 : $\tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}_0 = (\pi_{01} \cdots \pi_{0(C-1)})^T$, where π_{0C} is obtained by subtraction. For example, we may wish to test the null hypothesis that the frequency distribution of a phenotype follows Mendel's law for a trait determined by a single gene with dominant (A) and recessive (a) combinations AA , aa , and Aa that are expected to occur with specified probabilities $\boldsymbol{\pi}_0 = (0.25 \ 0.25 \ 0.5)^T$. The test of $H_0: \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}_0$ versus $H_1: \tilde{\boldsymbol{\pi}} \neq \tilde{\boldsymbol{\pi}}_0$ is then provided by the T^2 -like test of Wald (1943):

$$X_M^2 = N(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}_0)' \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\pi}_0)^{-1} (\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}_0), \quad (2.121)$$

(see Section A.7.1), where X^2 is asymptotically distributed as chi-square on $C - 1$ degrees of freedom χ_{C-1}^2 .

2.9.2 Pearson Chi-Square Test

The most widely known test for a multinomial distribution is the Pearson chi-square test (Pearson, 1900) that is a simple function of the observed frequencies in the various categories and the expected frequencies obtained under a specified null hypothesis. The simplest case is the test of the above hypothesis $H_0: \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}_0$, under which the

expected frequency in the j th category is $E(X_j) = E_j = N\pi_{0j}$. Then the Pearson chi-square test is provided by

$$X_{P,\pi_0}^2 = \sum_{j=1}^C \frac{(x_j - E_j)^2}{E_j}, \quad (2.122)$$

that is asymptotically distributed as χ_{C-1}^2 . Pearson's derivation is based on the demonstration that

$$\tilde{\Sigma}(\pi_0)^{-1} = \mathbf{D}_{\pi_0}^{-1} - \frac{\mathbf{J}\mathbf{J}'}{\pi_{0C}}, \quad (2.123)$$

where \mathbf{J} is a vector of ones of length $C - 1$ and $\mathbf{J}\mathbf{J}'$ is a matrix of ones. Substituting into (2.121) and expanding, it is then shown that

$$X_M^2 = N \sum_{j=1}^C \frac{(p_j - \pi_{0j})^2}{\pi_{0j}} = X_{P,\pi_0}^2. \quad (2.124)$$

Thus, the square of the large sample Z test for a single proportion in (2.5) equals the Pearson chi-square test (see Problem 2.14.3).

Pearson also extended the derivation to cases where the null hypothesis probabilities, and the expected frequencies, are estimated from the sample data under the null hypothesis to be tested. In the most general case, assume that the vector of probabilities is a function of a parameter vector $\theta = (\theta_1 \dots \theta_q)^T$, designated as $\pi(\theta)$, with q distinct parameters (i.e., eliminating those that are functions of the others). Then assume that the null hypothesis specifies that $\theta = \theta_0$. When θ_0 is prespecified, and thus $\pi_0 = \pi(\theta_0)$ is specified, the above test is distributed as χ_{C-1}^2 .

However, in most cases an estimate $\hat{\theta}_0$ may be obtained from the sample data, yielding estimated probabilities $\hat{\pi}_0 = \pi(\hat{\theta}_0)$ and estimated expected frequencies $\hat{E}(X_j) = \hat{E}_j = N\hat{\pi}_{0j}$ ($j = 1, \dots, C$). Then the Pearson chi-square test is provided by

$$X_P^2 = \sum_{j=1}^C \frac{(x_j - \hat{E}_j)^2}{\hat{E}_j}. \quad (2.125)$$

Taking a Taylor's expansion, Pearson showed that the distribution of X_P^2 converges in distribution to that of X_{P,π_0}^2 . However, he claimed that the degrees of freedom remained $C - 1$. Fisher (1922a) later showed that the degrees of freedom should allow for estimation of the q parameters $\hat{\theta}_0$, so that, asymptotically, $X_P^2 \xrightarrow{d} \chi_{C-q-1}^2$ on $C - q - 1$ degrees of freedom.

Example 2.14 The 2×2 Table

For the unconditional test of the difference in two proportions in (2.84), it is shown that the estimated expected frequencies under the null hypothesis of no difference between two independent groups can all be obtained from an estimate of the common probability of a positive outcome, $\hat{\pi}$, given the marginal frequencies in the rows and columns of the table. This is simply the estimate of $P(+)$, the complement $1 - \hat{\pi}$ being an estimate of $P(-)$. Likewise, let $\hat{\eta}$ designate the probability of membership in group 1, $1 - \hat{\eta}$ being an estimate of the probability of membership in group 2. Then

under the hypothesis of independence between group and response, the probabilities for each cell are the products of the corresponding marginal probabilities, so that the expected frequencies are

		Group		$N\hat{\pi}$
		1	2	
Response	+	$\hat{E}_{11} = N\hat{\pi}\hat{\eta}$	$\hat{E}_{12} = N\hat{\pi}(1 - \hat{\eta})$	$N\hat{\pi}$
	-	$\hat{E}_{21} = N(1 - \hat{\pi})\hat{\eta}$	$\hat{E}_{22} = N(1 - \hat{\pi})(1 - \hat{\eta})$	
		$N\hat{\eta}$	$N(1 - \hat{\eta})$	N

equal to those presented in Section 2.6.2. In terms of Fisher's general result, $\hat{\theta}_0 = (\hat{\pi}, \hat{\eta})$ consists of two elements, and thus the degrees of freedom is $4 - 2 - 1 = 1$.

2.9.3 Pearson Goodness-of-Fit Test

In some cases, as above, the null hypothesis specifies that there is no association among the two variables examined, such as group and response. In other cases one starts with an underlying model that is either prespecified, or in which the model parameters are estimated from the study data. Then the null hypothesis is that the observed data were generated under the assumed model from which the expected frequencies are obtained. The above test then provides a test of the goodness of fit of the model. Thus, the Pearson test is also commonly called the *Pearson goodness-of-fit test*. In fact, Pearson's test was originally derived to test whether a given member of the Pearson family of distributions provides a good fit to a set of data. An example will clarify.

Example 2.15 Recombination Fraction

Consider the characteristics of a crop of maize where the phenotypes can be defined from the combination of two characteristics: consistency (starch, A , vs. sugar, a) and color (green, B , vs. white, b), where the (A, B) phenotypes are dominant traits and the (a, b) traits are recessive. The probabilities of the four phenotypes (AB, Ab, aB, ab) of the two traits are functions of the recombination fraction $\sqrt{\theta}$ as

$$P(ab) = \pi_1 = \theta/4 \quad (2.126)$$

$$P(Ab) = P(aB) = \pi_2 = \pi_3 = (1 - \theta)/4$$

$$P(AB) = \pi_4 = (2 + \theta)/4,$$

where $0 < \theta < 1$. When the genes that determine each trait are in loci on different chromosomes, the traits are expected to segregate independently in which case $\theta = 0.25$ and the traits occur in the Mendelian ratios 1:3:3:9. Note that all parameters have been specified under H_0 .

Fisher (1925) presents a sample of $N = 3839$ observations with frequencies $x_1 = \#ab = 32$, $x_2 = \#Ab = 906$, $x_3 = \#aB = 904$, and $x_4 = \#AB = 1997$.

Thus, we wish to test the null hypothesis $H_0: \theta = \theta_0 = 0.25$. This yields expected frequencies $E_1 = 3839(1/16) = 239.94$, $E_2 = E_3 = 3839(3/16) = 719.81$, and $E_4 = 3839(9/16) = 2159.44$. Substituting into (2.122) for $C = 4$ yields a value of $X^2 = 287$, on $4 - 1 = 3$ df with $p < 0.0001$. Thus, the null hypothesis of independent segregation is rejected.

Fisher (1925) also shows that the maximum likelihood estimate of θ is $\hat{\theta} = 0.0357$. We might then wish to test the goodness of fit of the model using this value; or the null hypothesis that these two characteristics are located on the same chromosome with probabilities specified by the above recombination model with this estimate of θ . This yields expected frequencies of 34.27, 925.48, 925.48, and 1953.77 for the four cells. Substituting into (2.125) for $C = 4$ yields a goodness-of-fit test value $X_P^2 = 2.015$ on $4 - 1 - 1 = 2$ df with $p = 0.37$. Thus, we fail to reject the hypothesis that this model fits the observed data. Of course, this does not prove that this model is the true model, only that we cannot reject the hypothesis that it is the true model.

2.9.4 Logits

For a binary variable with a single "positive" category of primary interest, other summary measures include the odds, and its log (or the logit). For a polychotomous variable, owing to the greater number of categories and to possibly ordered categories, various functions of the sample frequencies or proportions are often employed in an analysis, especially for ordinal data, where there is an underlying order to the categories, such as normal, mildly abnormal, moderately abnormal, or severely abnormal, as often used in pathology.

For a classification with C categories there are $C - 1$ degrees of freedom. This implies that the information in the data can be described using $C - 1$ distinct summary statistics, such as the log odds or logits comparing each category to a reference category. For example, for the recombination data it is typical to use the rarer phenotype to represent the reference category, so that the categories can be labeled arbitrarily as 1 = ab , 2 = Ba , 3 = Ab , and 4 = AB . Then the *pairwise* or *generalized logits* are defined as above for the case of a binary variable as in (2.8), such as

$$\theta_j = \log[\pi_j/\pi_1], \quad j = 2, \dots, C \quad (2.127)$$

with covariance matrix as shown in Example A.4. The sample estimates, computed from the proportions within each category, then are distributed as multivariate normal with a covariance matrix of rank $C - 1$ (i.e., such that the distribution is not degenerate).

For a specific ordering of the categories, such as the above ordering of the phenotypes, or for a variable with inherently ordered categories, another approach is to use *continuing-ratio logits*, defined as the log odds of a given category relative to all categories that precede it. Let $\pi_{j+} = \sum_{k=1}^j \pi_k$. This yields the continuation logits

$$\theta_j^+ = \log[\pi_j/\pi_{(j-1)+}], \quad j = 2, \dots, C. \quad (2.128)$$

As in Example A.4, it follows (see Exercise 2.14.4) that the variances of the sample estimates are

$$V(\hat{\theta}_j^+) = N^{-1} \left[\pi_{(j-1)^+}^{-1} + \pi_j^{-1} \right], \quad j = 2, \dots, C \quad (2.129)$$

with covariances $\text{Cov}(\hat{\theta}_j^+, \hat{\theta}_k^+) = 0$, for $2 \leq j < k \leq C$.

Various additional types of logits could also be employed. Among these, the *cumulative logits* assess the odds of falling in categories above (and including) a cutpoint versus categories below. Such logits are the basis for the *proportional odds model* that is a generalization of the logistic regression model for a binary outcome to a model for an ordinal outcome. The proportional odds model is described in Chapter 7.

Example 2.16 Recombination Fraction

Again consider the above data from Fisher (1925). The following table presents the proportions within each category, the generalized odds and logits relative to the first category as the reference, their variance and 95% confidence limits, and the resulting asymmetric confidence limits on the odds

j	p_j	$\widehat{\text{Odds}}_j$	$\widehat{\theta}_j$	$\widehat{V}(\widehat{\theta}_j)$	$\widehat{\theta}_{jL}$	$\widehat{\theta}_{jU}$	$\widehat{\text{Odds}}_{jL}$	$\widehat{\text{Odds}}_{jU}$
1: <i>ab</i>	0.0083	1.0	—	—	—	—	—	—
2: <i>Ab</i>	0.2360	28.3	3.34	0.0324	2.99	3.70	19.9	40.3
3: <i>Ba</i>	0.2355	28.3	3.34	0.0324	2.99	3.70	19.9	40.2
4: <i>AB</i>	0.5202	62.4	4.13	0.0318	3.78	4.48	44.0	88.5

These odds depart substantially from the 1, 3, 3, and 9 that would be expected under Mendelian inheritance.

Since these categories are not truly ordinal, the continuation logits would not be meaningful. These are illustrated in Example 2.7.

2.10 TWO INDEPENDENT GROUPS WITH POLYCHOTOMOUS RESPONSE

2.10.1 Large Sample Test of Proportions

Now consider the case of two independent groups with n_1 and n_2 observations, respectively, of a polychotomous outcome that we wish to compare. Denote the vector of frequencies within each group as $\mathbf{x}_i = (x_{1(i)} \cdots x_{C(i)})^T$ and the corresponding vector of proportions within the C categories as $\mathbf{p}_i = (p_{1(i)} \cdots p_{C(i)})^T$, each with expectation $\boldsymbol{\pi}_i = (\pi_{1(i)} \cdots \pi_{C(i)})^T$ where $p_{j(i)}[\pi_{j(i)}]$ is the conditional proportion [probability] of falling in the j th category in the i th group. To allow for the linear dependence among these proportions and probabilities, let $\tilde{\mathbf{p}}_i$, $\tilde{\boldsymbol{\pi}}_i$, and $\tilde{\Sigma}(\boldsymbol{\pi}_i)$ denote the corresponding vectors/matrix for the $C - 1$ elements. Since the two groups are independent, then the difference in the vectors of proportions, $\mathbf{d} = \tilde{\mathbf{p}}_1 - \tilde{\mathbf{p}}_2$, is distributed as multivariate normal with expectation $\Delta = \tilde{\boldsymbol{\pi}}_1 - \tilde{\boldsymbol{\pi}}_2$ and covariance

matrix

$$\Sigma_{\Delta} = \frac{\tilde{\Sigma}(\boldsymbol{\pi}_1)}{n_1} + \frac{\tilde{\Sigma}(\boldsymbol{\pi}_2)}{n_2} \quad (2.130)$$

with $\Sigma(\boldsymbol{\pi}_i)$ as shown in (A.26) with respect to the probabilities within each group.

Under the null hypothesis $H_0: \Delta = \Delta_0 = \mathbf{0}$, then $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = \boldsymbol{\pi}$ that is consistently estimated by the pooled vector of proportions

$$\mathbf{p} = (n_1 \mathbf{p}_1 + n_2 \mathbf{p}_2)/N = (\mathbf{x}_1 + \mathbf{x}_2)/N \quad (2.131)$$

with the $C - 1$ subvector $\tilde{\mathbf{p}}$. Under H_0 ,

$$\Sigma_{\Delta_0} = \tilde{\Sigma}(\boldsymbol{\pi}) \frac{N}{n_1 n_2} \quad (2.132)$$

that is consistently estimated as

$$\hat{\Sigma}_{\Delta_0} = \tilde{\Sigma}(\mathbf{p}) \frac{N}{n_1 n_2}. \quad (2.133)$$

The test of H_0 versus $H_1: \Delta \neq \mathbf{0}$ is then provided by the T^2 -like test of Wald (1943)

$$X^2 = \mathbf{d}' \hat{\Sigma}_{\Delta_0}^{-1} \mathbf{d} \quad (2.134)$$

that is asymptotically distributed as χ^2_{C-1} (see Section A.7.1).

2.10.2 The Pearson Contingency Chi-Square Test

Algebraically, it can again be shown (see Problem 2.15.2) that the above test for the difference in two vectors of proportions is equal to a Pearson test of the form

$$X_P^2 = \sum_{i=1}^2 \sum_{j=1}^C \frac{(x_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}. \quad (2.135)$$

As was the case for a binary outcome in a 2×2 table, for a polychotomous outcome, the hypothesis of a common vector of response probabilities is equivalent to the hypothesis of independence between the row (group) and column (category) marginal classifications. The vector of estimated marginal probabilities of each response category is $\hat{\boldsymbol{\pi}} = \mathbf{p} = (\hat{\pi}_1 \cdots \hat{\pi}_C)^T$ and the estimated probabilities of membership in the two groups are $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \hat{\eta}_2)$ where $\hat{\eta}_1 = n_1/N$ and $\hat{\eta}_2 = 1 - \hat{\eta}_1$. Then under the independence hypothesis, the expected frequencies are $\hat{E}_{ij} = N \hat{\eta}_i \hat{\pi}_j$.

From Fisher's result, the minimum estimated parameters of $q = C$ elements are $\hat{\boldsymbol{\theta}}_0 = (\hat{\pi}_1 \cdots \hat{\pi}_{C-1} \hat{\eta}_1)$ and thus the degrees of freedom are $2C - C - 1 = C - 1$ df. Thus, asymptotically $X_P^2 \xrightarrow{d} \chi^2_{C-1}$ on $C - 1$ degrees of freedom.

2.10.3 Odds Ratios

Note that the test of equality of proportions between groups is also a test of the equality of the generalized or pairwise logits between groups, or that the corresponding odds ratios among groups are equal to 1. Let $\theta_j = \log(\pi_{j(1)}/\pi_{1(1)}) - \log(\pi_{j(2)}/\pi_{1(2)})$ denote the log odds ratio for the j th category relative to the first reference category ($j = 2, \dots, C$). Then from Example A.4, it is readily shown (see Problem 2.15.3) that the sample estimate $\hat{\theta}_j = \log(p_{j(1)}/p_{1(1)}) - \log(p_{j(2)}/p_{1(2)})$ is asymptotically normally distributed with mean θ_j and covariance matrix with elements

$$V(\theta_j) = \frac{1}{n_1} \left[\frac{1}{\pi_{1(1)}} + \frac{1}{\pi_{j(1)}} \right] + \frac{1}{n_2} \left[\frac{1}{\pi_{1(2)}} + \frac{1}{\pi_{j(2)}} \right], \quad j = 2, \dots, C \quad (2.136)$$

$$Cov(\theta_j, \theta_k) = \frac{1}{n_1 \pi_{1(1)}} + \frac{1}{n_2 \pi_{1(2)}}, \quad 1 < j < k \leq C.$$

These quantities are consistently estimated when evaluated using the sample proportions. When exponentiated, the confidence limits on θ_j then provide asymmetric confidence limits on the generalized odds ratios.

The continuation-ratio logits and the corresponding odds ratios could also be used to describe the difference between the two groups. However, this adds little to the generalized logits.

The fact that the continuation-ratio logits have covariance zero, as shown in (2.129), also implies that the chi-square test for the $2 \times C$ table on $C-1$ *df* can be expressed as the sum of $C-1$ independent 1 *df* tests, each corresponding to a continuation logit. For the j th category ($j > 2$), the 1 *df* test is obtained from the 2×2 table constructed by collapsing the $j-1$ preceding categories; e.g., for $j = 3$, a table comparing category 3 versus categories 1 + 2. This is a special case of partitioning a $R \times C$ contingency table into $(R-1)(C-1)$ separate independent 2×2 tables such that the Pearson chi-square test on $(R-1)(C-1)$ *df* can be expressed as the sum of the 1 *df* chi-square statistics for each of the separate 2×2 tables. Such partitionings were described by Irwin (1949) and Lancaster (1949, 1950), see also Kimball (1954).

2.10.4 Rank Tests: Cochran-Mantel-Haenszel Mean Scores Test

For an inherently ordered variable, the statistical analysis will often be based on numerical scores assigned to the C categories. The simplest are the *table scores* consisting of the numerical values assigned to each category. Unless the categories represent a true count, the table scores are arbitrary. For example, if there are three ordered nonnumerical categories, they could be assigned any set of three ordered values such as $\{0, 1, 2\}$ or $\{1, 3, 5\}$. Since the two sets of numbers are not proportional to a constant, the results of an analysis using these two sets of scores could differ.

Other possible scores are based on the ranks of the observations. For a categorical variable with C -ordered categories, there will be many tied values, and thus many tied ranks. Let $\mathbf{x} = (x_1 \cdots x_C)^T$ denote the vector of marginal frequencies within each category. The rank scores, say r_j , assigned to each category $j = 1, \dots, C$ are

$$r_1 = \frac{\sum_{i=1}^{x_1} i}{x_1} = \frac{x_1(x_1 + 1)/2}{x_1} = \frac{x_1 + 1}{2}, \quad (2.137)$$

$$r_j = \sum_{k=1}^{j-1} x_k + \frac{x_j + 1}{2}, \quad j > 1.$$

Note that the rank score for the j th category is the average of all the "untied" possible ranks that would be assigned if there were x_j distinct values. The resulting score is based on the fact that $\sum_{i=1}^a i = a(a + 1)/2$. These are also called the *midranks*.

The midranks are a function of the sample size. Dividing by N yields the fractional ranks, also called RIDITS (Bross, 1958), or dividing by $N + 1$ yields the modified ridits based on the order statistics from a unit uniform distribution (van Elteren, 1960; Lehmann, 1975). The term *ridit* comes from the term "relative to an identified distribution," with the suffix "it" analogous to logit. These rank scores provide generalizations of nonparametric analyses, such as the Wilcoxon rank sum test (Wilcoxon, 1945), to ordered categorical data.

There are many other possible score functions that could be applied. Hájek and Šídák (1967) provide a review and describe the optimal score generating function under a specific alternative hypothesis, such as a location shift under a specific continuous distribution. For example, the Wilcoxon test based on rank scores is optimal for a location shift under a logistic distribution. Other scores are optimal under other alternatives.

Cochran (1954) and Mantel and Haenszel (1959) each described a similar 1 *df* test that is termed the *Cochran-Mantel-Haenszel mean scores test*. The test statistic is a linear combination of the observed proportions within each group using the scores as the weights, in contrast to the Wilcoxon and other linear rank tests, where the statistic is a linear function of the scores. Let $\mathbf{s} = (s_1 \cdots s_C)^T$ denote the vector of numerical scores assigned to the C ordered categories, such as rank scores. Then a simple summary measure is the mean score, computed as

$$\bar{s} = \frac{\sum_{j=1}^C s_j x_j}{N} = \frac{\mathbf{s}' \mathbf{x}}{N} = \mathbf{s}' \mathbf{p}. \quad (2.138)$$

Note that this test is computed as a simple linear combination of the frequencies using the vector \mathbf{s}/N . Since the vector \mathbf{x} is distributed as multivariate normal, the \bar{s} is distributed as normal with expectation and large sample variance

$$E(\bar{s}) = \mathbf{s}' \boldsymbol{\pi} \quad (2.139)$$

$$V(\bar{s}) = \mathbf{s}' \boldsymbol{\Sigma}(\boldsymbol{\pi}) \mathbf{s} / N = \mathbf{s}' [\mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi} \boldsymbol{\pi}'] \mathbf{s} / N.$$

For two independent groups, a test of $H_0: \mathbf{s}'\boldsymbol{\pi}_1 = \mathbf{s}'\boldsymbol{\pi}_2 = \mathbf{s}'\boldsymbol{\pi}$ is provided by

$$X_s^2 = \frac{[\mathbf{s}'(\mathbf{p}_1 - \mathbf{p}_2)]^2}{\mathbf{s}'\boldsymbol{\Sigma}(\mathbf{p})\mathbf{s} \left[\frac{1}{n_1} + \frac{1}{n_2} \right]} = \frac{\bar{d}^2}{V(\bar{d})}, \quad (2.140)$$

where $\hat{\boldsymbol{\pi}} = \mathbf{p}$ and $\bar{d} = \bar{s}_1 - \bar{s}_2$.

For the case of two groups, this mean score test provides the same result using the rank, rudit or modridit scores, since all are proportional. This test is equivalent to the Wilcoxon rank sum test for two groups, without a continuity correction.

Example 2.17 *Retinopathy in the DCCT*

As described in Section 1.5, an objective of the Diabetes Control and Complications Trial (DCCT, 1993, 1995d) was to assess the effects of intensive versus conventional management of diabetes glucose control on the risks of diabetic retinopathy (microvascular abnormalities of the retina). Owing to staggered entry, the following analyses are based on the cohort of subjects evaluated at five years of follow-up. The analysis is restricted to those who were enrolled into the secondary cohort of subjects with 1 to 15 years' duration and minimal to mild nonproliferative diabetic retinopathy on entry.

Diabetic retinopathy (DR) severity for each subject was graded on a multiple-step scale that allowed definition of outcomes in terms of the number of steps change as well as landmarks of severity. The minimal change of interest was progression of at least three steps along the scale from the level at baseline (sustained at two successive visits). More severe outcomes of interest were progression to the level of severe nonproliferative retinopathy (NPDR), or to a worse level that required panretinal photocoagulation (laser surgery) to preserve vision. Thus, the status of each subject at five years can be categorized as not having shown sustained progression ≥ 3 steps, or with ≥ 3 steps sustained progression (but not to severe NPDR), severe NPDR (but without the need for laser surgery), or laser surgery having been required. As defined, these categories are mutually exclusive. The following table presents the results:

Retinopathy	Group		Total
	Intensive	Conventional	
0: < 3 steps	307 (89.5%)	260 (79.3%)	567 (84.5%)
1: 3+ Steps	19 (5.5%)	33 (10.1%)	52 (7.8%)
2: Severe NPDR	12 (3.5%)	18 (5.5%)	30 (4.5%)
3: Laser Rx	5 (1.5%)	17 (5.2%)	22 (3.3%)
Total	343	328	671

Treating this as a 4×2 table yields a Pearson chi-square test value of $X_P^2 = 15.1$ on 3 df with $p = 0.0017$. Although highly significant, this test is less powerful because it ignores the ordinal nature of the outcome variable.

Table 2.7 Cochran-Mantel-Haenszel rank statistics.

SAS PROC FREQ Analysis of 2xC tables
 DCCT Retinopathy Scores, Example 2.17
 Table of GROUP by retlevel5y

GROUP(Treatment Group (Exp,Std)) retlevel5y		Frequency	0	1	2	3	Total
EXPERIMENTAL		307	19	12	5	343	
STANDARD		260	33	18	17	328	
Total		567	52	30	22	671	

Summary Statistics for GROUP by retlevel5y						
Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)						
Statistic	Alternative Hypothesis	DF	Value	Prob		
1	Nonzero Correlation	1	13.7954	0.0002		
2	Row Mean Scores Differ	1	13.7954	0.0002		
3	General Association	3	15.0604	0.0018		

A mean score test using rank (or ridit or modified ridit) scores yields $X_s^2 = 13.8$ on 1 *df* with $p = 0.00021$, the greater power being due to the ordinal data that allows a 1 *df* test, resulting in a smaller *p*-value.

The following table shows the generalized odds ratios relative to step 0 as the reference within the two groups and the asymmetric confidence limits based on the log odds ratio.

Categories	Odds Ratio	95% Confidence Limits	
		Lower	Upper
2:1	2.05	1.14	3.69
3:1	1.77	0.84	3.75
4:1	4.01	1.46	11.03

This shows an approximately two-fold greater odds of levels 2 or 3 versus level 1 with conventional versus intensive therapy, and a four-fold greater odds of level 4.

Since these data are ordinal in nature, it might also be of interest to assess the continuation-ratio odds ratios. The results are as follows:

Categories	Odds	95% Confidence Limits	
	Ratio	Lower	Upper
2:1	2.05	1.14	3.69
3:2 ⁺	1.67	0.79	3.52
4:3 ⁺	3.70	1.35	10.13

The above tests of significance can be provided by the SAS FREQ procedure. The following statements provide the test using rank scores:

```
proc freq; table group*retlevel5y
    / nopercent nocol cmh scores=rank;
```

and produce the output shown in Table 2.7 (edited). The test X_s^2 in (2.140) employs the unconditional multinomial covariance matrix, whereas SAS uses the conditional covariance matrix for the multinomial. Thus, the value produced by SAS equals $X_s^2(N - 1)/N$, resulting in a slightly smaller test value of 13.80 that is significant at $p = 0.00021$. Since there are only two groups, this test value is equivalent to the test of nonzero correlation.

The test of general association is the CMH equivalent of the Pearson test of independence between the row and column classifications but is not relevant for these data because the response variable is ordinal.

The program does not have an option to compute the generalized or cumulative odds ratios. These were computed using the author's program *retinopdata.sas*.

2.11 MULTIPLE INDEPENDENT GROUPS

2.11.1 The Pearson Test

The developments above also generalize to the case of $R \geq 2$ multiple independent groups with a C -category nominal or ordinal response or dependent measure, generating a $R \times C$ contingency table. Let x_{ij} denote the joint frequency within the i th row and j th column with marginal frequencies $x_{i\bullet}$ among the rows and $x_{\bullet j}$ among the columns, $i = 1, \dots, R$; $j = 1, \dots, C$; where the " \bullet " represents summation over the respective index. The corresponding joint probabilities are designated as π_{ij} , with marginal probabilities $\pi_{i\bullet}$ and $\pi_{\bullet j}$ for the i th group (row) and j th response (column) categories, respectively.

The R row categories could be defined from one polychotomous variable, such as ethnicity, and the column categories likewise defined from another variable, such as grades of illness severity, when the two are measured in a single cross-sectional sample from a population. Then the frequencies within the $R \times C$ table are distributed as multinomial with RC categories defined from combinations of the two variables,

and the probability of the table is the multinomial probability

$$P(x_{11}, \dots, x_{RC}) = \frac{N!}{\prod_{i=1}^R \prod_{j=1}^C x_{ij}!} \prod_{i=1}^R \prod_{j=1}^C \pi_{ij}^{x_{ij}}. \quad (2.141)$$

The null hypothesis of statistical independence between the row and column classification factors then states $H_0: \pi_{ij} = \pi_{i\bullet} \pi_{\bullet j}$. The alternative hypothesis, $H_1: \pi_{ij} \neq \pi_{i\bullet} \pi_{\bullet j}$, specifies that there is some degree of association between the row and column factors. Under H_0 , $E(x_{ij}) = E_{ij} = N\pi_{i\bullet} \pi_{\bullet j}$. The sample estimates of the marginal probabilities $\hat{\pi}_{i\bullet} = n_{i\bullet}/N$ and $\hat{\pi}_{\bullet j} = n_{\bullet j}/N$ yield the estimated expected frequencies $\hat{E}_{ij} = n_{i\bullet} n_{\bullet j}/N$.

Alternatively, the row categories could represent separate samples of subjects from R different populations, such as a study in which multiple treatments are compared. Conditioning on the sample sizes within each group, as would be appropriate for such a comparative study, the frequencies within each group are distributed as multinomial with conditional probabilities $\pi_i = (\pi_{1(i)} \dots \pi_{C(i)})^T$, and the probability of a given table is provided by the product of the R separate multinomial probabilities with separate vectors of parameters $\{\pi_i\}$, or the product multinomial likelihood

$$P(x_{1(1)}, \dots, x_{C(R)}) = \prod_{i=1}^R \frac{n_{i\bullet}!}{\prod_{j=1}^C x_{j(i)}!} \prod_{j=1}^C \pi_{j(i)}^{x_{j(i)}}, \quad (2.142)$$

where $x_{j(i)} = x_{ij}$. As for the case of two independent groups in the preceding section, the null hypothesis of homogeneity specifies that the vector of parameters is the same for all groups, or $H_0: \pi_1 = \dots = \pi_R = \pi = (\pi_{\bullet 1} \dots \pi_{\bullet C})^T$. The vector of estimated marginal probabilities is provided by $\hat{\pi} = (\hat{\pi}_{\bullet 1} \dots \hat{\pi}_{\bullet C})^T = (\hat{p}_{\bullet 1} \dots \hat{p}_{\bullet C})^T$. Likewise, the estimated probabilities of membership in the R groups are $\hat{\eta} = (\hat{\eta}_1 \dots \hat{\eta}_R)$, where $\hat{\eta}_i = \hat{\pi}_{i\bullet}$. Under the hypothesis of homogeneity, the expected frequencies are $\{\hat{E}_{ij}\} = N\hat{\eta}_i \hat{\pi}_j = n_{i\bullet} n_{\bullet j}/N$ that equal those derived under the hypothesis of independence above.

Then the Pearson test of either independence between rows and columns, or of homogeneity among the R groups, is obtained as

$$X_P^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(x_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}. \quad (2.143)$$

Since the number of distinct estimated parameters is $(R - 1) + (C - 1)$, the test is asymptotically distributed as chi-square on $RC - (R - 1) - (C - 1) - 1 = (R - 1)(C - 1)$ df, or $X_P^2 \stackrel{d}{\approx} \chi_{(R-1)(C-1)}^2$.

The Cochran-Mantel-Haenszel test of general association is asymptotically equivalent to the Pearson test, but it is computed using the conditional covariance matrix so that it equals $X_P^2(N - 1)/N$.

2.11.2 Measures of Association

A variety of measures have been proposed to describe the strength of the association between the two categorical variables in an $R \times C$ table. Cramér's V (Cramér, 1946) is appropriate for a table constructed from two categorical variables, or for any table where the Pearson contingency statistic on $(R - 1)(C - 1)$ df is employed. It is readily shown that the maximum possible value of X_P^2 is $\min[R - 1, C - 1]N$, so that

$$V = \sqrt{\frac{X_P^2}{\min[R - 1, C - 1]N}}, \quad (2.144)$$

where $0 \leq V \leq 1$. Thus, V is analogous to a correlation coefficient and V^2 to a measure of R^2 . However, it is not a direct R^2 measure in the sense of the fraction of explained variation on some scale.

As noted in Section 2.7, the uncertainty coefficient $U(C|R)$ for the column classification as a function of the row classification for a 2×2 table equals the entropy R^2 described in Chapter 7 that is a measure of the strength of association in a logistic regression model with a binary dependent variable and a binary independent variable. The entropy R^2 is computed as the percentage reduction in the $-2 \log(L)$ for the logistic model attributable to adding group to the model (L being the likelihood). For a $R \times C$ table, the coefficient $U(C|R)$ also equals the percentage reduction in the $-2 \log(L)$ in a polychotomous logit model for the C categories as a function of the R groups. Thus, $U(C|R)$ is a direct R^2 measure.

2.11.3 Logits

As shown in Section 2.10.3 for the $2 \times C$ table, it is possible to decompose the test value for the $R \times C$ table into the sum of $(R - 1)(C - 1)$ statistically independent 1 df chi-square values, each based on a partitioning of the table into a 2×2 subtable for which an odds ratio could be computed. Among the possible partitions are those formed by successive collapsing of rows and columns as used to construct the continuation-ratio logits and odds ratios. However, these are rarely of practical use in describing the differences among groups.

However, odds ratios, or other summary measures, such as a relative risk, may be helpful when there is a binary response ($C = 2$). In this case, the difference among the R groups is represented by the vector of conditional proportions "positive" among the groups in the form of $R(R - 1)/2$ pairwise comparisons among the groups, or $(R - 1)$ comparisons against a reference or a common control group. Such multiple tests, however, increase the type I error probability for the set of such tests.

2.11.4 Multiple Tests

The chi-square test for the $R \times C$ table (2.143) provides a test of whether there is departure from the hypotheses of independence between the row and column

categories, or of homogeneity of the C response probabilities among the R groups. Although helpful in determining whether any such difference exists, the overall test cannot identify the nature of the pattern of differences between groups. For this purpose, some form of multiple tests may be performed, in which case the probability of a type I error may be inflated.

Let K refer to the number of distinct tests to be conducted, each with a defined null hypothesis H_{0k} , for $k = 1, \dots, K$. If the R levels are nominal categories (groups), all $K = R(R - 1)/2$ pairwise comparisons might be performed, each using a chi-square test comparing a pair of groups on $C - 1$ *df*. For categories $1 \leq i < \ell \leq R$, the index is obtained as $k = (\ell - i) + (i - 1)(C - i/2)$.

Alternatively, if one group is a reference, then the $R - 1$ remaining groups might each be compared to this reference using a test on $C - 1$ *df*. The latter would apply, for example, when $R - 1$ treatments are compared to a common control group. In this case $k = (i - 1)$ for the i th group versus control and $K = R - 1$.

The null hypothesis for the k th test specifies that the row and column categories in that $2 \times C$ table are independent, or that the vectors of probabilities are equal for the two groups being compared; i.e., H_{0k} : $\pi_i = \pi_\ell$ for the i th versus the ℓ th groups, $k = 1, \dots, K$. The joint null hypothesis H_0 then is the set of K such pairwise hypotheses. If all K tests were conducted using the nominal $p \leq 0.05$ to declare statistical significance, then the probability of a type I error α for the set of K tests would be inflated. Thus, a more stringent criterion is required to declare significance so as to protect against this inflation.

The Bonferroni adjustment is the simplest approach to adjusting the significance level for the K tests. This is based on the Bonferroni inequality using the first term in the Bonferroni expansion for the probability of a joint "event" such that

$$P(\text{at least one of } K \text{ tests significant at level } \alpha' | H_0) \leq \alpha' K.$$

Therefore, to achieve a total Type I error probability no greater than the desired level α , we could use the significance level $\alpha' = \alpha/K$ for each of the K separate tests. Likewise, to construct K confidence intervals with the desired coverage probabilities no less than the desired confidence level $1 - \alpha$, the K confidence limits could be constructed using the confidence level $1 - \alpha'$. For example, for $K = 4$ we would use $\alpha' = 0.05/4 = 0.0125$ to conduct each of the tests of significance and to compute the corresponding confidence limits on the difference parameter for each comparison.

The Bonferroni inequality, however, is unduly conservative, thus needlessly sacrificing power. A variety of less conservative multiple test procedures have been proposed, such as the step-down procedure of Holm (1979), the step-up procedure of Hochberg (1988), and that of Hommel (1988). These procedures are members of the general family of closed test procedures of Marcus et al. (1976) that provide strict control of the type I error probability for the set (Hommel, 1988; Liu, 1996) under specific assumptions.

These procedures conduct the multiple tests after ordering the K test *p*-values from smallest (say, p^1) to largest (say, p^K) and then applying less restrictive significance levels than imposed by the Bonferroni procedure to the second, third, and so on, ordered test. The critical *p*-values employed by the Holm and Hochberg procedures

are as follows:

Ordered p -value	$p^{(1)}$	$p^{(2)}$	\dots	$p^{(k)}$	\dots	$p^{(K-1)}$	$p^{(K)}$
Critical p -value	$\frac{\alpha}{K}$	$\frac{\alpha}{K-1}$	\dots	$\frac{\alpha}{K-k+1}$	\dots	$\frac{\alpha}{2}$	α

The Holm procedure is also called *step-down Bonferroni* because it requires a significance level of α/K for the first test (maximum statistic, minimum p -value), $\alpha/(K-1)$ for the second smallest p -value, and so on. When any one test is not significant, the procedure stops and all further tests are also declared nonsignificant.

The Hochberg procedure employs the same critical values, but the order of testing is reversed. If the maximal p -value is $> \alpha$, then the next-largest comparison is tested. If its p -value is $> \alpha/2$, then again the next is tested with criterion $\alpha/3$ and so on. When any one of these satisfies significance; i.e., if $p_k \leq \alpha/(K-k+1)$ for the k th smallest p -value, then that and all remaining tests are declared to be statistically significant. The Hommel (1988) procedure is more complicated and the details are not described herein.

These and other multiple testing procedures are readily implemented using the SAS PROC MULTTEST that provides adjusted p -values illustrated in the following Example 2.18. For the Bonferroni procedure, the adjusted p -value for the k th test is simply $\tilde{p}^k = Kp^k$. For the Holm procedure, the adjusted p -values are computed as

$$\tilde{p}^k = \begin{cases} Kp^1, & k = 1 \\ \max[\tilde{p}^{k-1}, p^k(K-k+1)], & k = 2, \dots, K, \end{cases} \quad (2.145)$$

where any adjusted p -value > 1 is set to 1. For the Hochberg procedure,

$$\tilde{p}^k = \begin{cases} p^K, & k = K \\ \min[\tilde{p}^{k+1}, p^k(K-k+1)], & k = (K-1), \dots, 2 \end{cases} \quad (2.146)$$

The \tilde{p}^k for the Hommel procedure are described by Wright (1992). If any $\tilde{p}^k \leq \alpha$, then that test is significant adjusted for the set of K tests.

Liu (1996) shows that the Hommel procedure dominates that of Hochberg, that dominates that of Holm, that dominates that of Bonferroni, meaning that an adjusted p -value using the Hommel procedure is \leq that of Hochberg that is \leq that of Holm. However, the Hochberg and Hommel procedures were strictly proven to control α only for the case of independent tests. While simulations have shown that these procedures provide control of α in general for correlated tests, as is the case for pairwise comparisons, and in fact tend to be conservative, it is possible that an unusual situation may arise where the level of α is not controlled.

Example 2.18 *Ethnicity and Hypertension in the Diabetes Prevention Program*

The Diabetes Prevention Program studied approximately 4000 individuals with impaired glucose tolerance. Ethnicity and hypertension are known risk factors for type 2 diabetes. Thus, the association of these two factors was evaluated (DPP, 2002)

Table 2.8 Pairwise p -values comparing ethnicities in the DPP. Raw p -values and rank order above the diagonal. Hochberg adjusted p -values below the diagonal.

	White	African American	Hispanic	Other
White	—	< 0.001 ³	0.0038 ⁴	0.4788 ⁶
African-Am.	< 0.001	—	< 0.001 ²	< 0.001 ¹
Hispanic	0.0114	< 0.001	—	0.1593 ⁵
Other	0.4788	0.0013	0.3187	—

with the following frequency table:

Hypertension

	White	African American	Hispanic	Other
Yes #	579	275	131	87
%	27.4%	36.6%	21.5%	25.5%
Total	2117	752	609	341

The Pearson chi-square test of independence yields $X_P^2 = 41.5$ on 3 df with $p < 0.0001$.

These data suggest that African-American subjects may have a higher prevalence of hypertension, and Hispanic subjects a lower prevalence, than other ethnicities. Pairwise comparisons then assess whether these differences are statistically significant with a joint type I error probability no greater than 0.05.

Table 2.8 presents the nominal p -values for all possible pairwise comparisons in the upper right elements above the diagonal. The superscript is the rank order based on the numerical test value, so that tests with $p < 0.001$ can be differentiated.

Consider the Hochberg procedure. Starting from the largest p -value, the first comparison that meets the criterion for significance is 0.0038 ($k = 4$), for which the corresponding critical value for significance at the 0.05 level is $0.05/(6 - 4 + 1) = 0.0167$. Thus, it and all other comparisons with a lower rank order are significant at this level. The resulting adjusted p -values are shown below the diagonal in the table. For example, since the critical value for $p^4 = 0.0038$ is $0.05/3$, then the Hochberg adjusted p -value for this test is $0.0038 \times 3 = 0.0114$. By this criterion, adjusted for six tests, African Americans had significantly higher risk than the other groups, and Hispanics had significantly lower risk than Caucasians, with an adjusted $p \leq 0.05$.

Adjusted p -values could also be computed using the SAS procedure MULTTEST. The data set must contain a variable named "raw-p" that provides the individual p -values, along with a variable "test" that is the identifier for each p -value. The following statement would provide the Holm (step-down Bonferroni), Hochberg, and Hommel adjusted p -values.

```
proc multtest pdata=pvals bon holm hoc hom;
```

For this example, it turns out that the three procedures provide the same adjusted p -values, although that will not always be the case.

Example 2.19 Hypothetical p -Values

Consider the application of PROC MULTTEST to the set of six hypothetical p -values: 0.00135, 0.00852, 0.06371, 0.06371, 0.11208, and 0.32944 that yields the following adjusted p -values:

Test	Stepdown				
	Raw	Bonferroni	Bonferroni	Hochberg	Hommel
1	0.0014	0.0081	0.0081	0.0081	0.0081
2	0.0085	0.0511	0.0426	0.0426	0.0426
3	0.0637	0.3823	0.2548	0.1911	0.1681
4	0.0637	0.3823	0.2548	0.1911	0.1681
5	0.1121	0.6725	0.2548	0.2242	0.2242
6	0.3294	1.0000	0.3294	0.3294	0.3294

This illustrates that different adjustments can indeed yield different adjusted p -values.

2.11.5 Rank and Correlation Tests

When the independent variable is nominal with R categories (groups) and the dependent variable is ordinal with C ordered categories, the null hypothesis of interest $H_0: \pi_1 = \dots = \pi_R = \pi$ can be tested using a generalization of (2.140) using the mean of the scores among the C categories within each of the R groups. The reader is referred to Agresti (1990) for the computational details. This results in a test statistic that is distributed as chi-square on $R - 1$ df . With rank scores this test is equivalent to the Kruskal and Wallis (1952) generalization of the Wilcoxon rank sum test to $R > 2$ groups.

When the independent variable is also ordinal with R ordered categories, the hypothesis of interest is of no correlation among the ordered categories. The Pearson correlation between the two sets of scores then provides the basis for a 1 df test. When used with rank or ridit scores, this provides the Spearman rank correlation (Spearman, 1904) and its associated test of significance. The Pearson correlation would be appropriate when the numerical values for each variable had intrinsic quantitative meaning, such as a count.

Example 2.20 Coronary Heart Disease in the Framingham Study

Cornfield (1962) described the association between ordered categories of blood pressure and cholesterol with the risk of coronary heart disease over six years of follow-up in the Framingham study. Here we use these data to assess the association between the level of cholesterol with the level of blood pressure in the cohort. The following

table presents the observed frequencies in the cross classification.

Serum Cholesterol (mg/dL)	Systolic Blood Pressure (mmHg)				Total
	<127	127–146	147–166	167+	
<200	119	124	50	26	319
200–219	88	100	43	23	254
220–259	127	220	74	49	470
260+	74	111	57	44	286
<i>Total</i>	408	555	224	142	1329

(reproduced with permission). In this case, the relevant test is the test of nonzero correlation between the two categories using rank scores because no unique value can be associated with each category. SAS PROC FREQ with the *cmh* option and rank scores yields $X^2 = 16.5284$ on 1 *df* with $p < 0.0001$.

2.11.6 The Cochran-Armitage Test for Trend

Cochran (1954a) and Armitage (1955) described a test of linear trend in the proportions positive among R ordered categories (groups) with a binary response ($C = 2$). Within each group of sample size n_i , let s_i denote the corresponding score and p_i the proportion positive with expectation π_i , $i = 1, \dots, R$. We then assume that the π_i are a linear function of the s_i of the familiar form $\pi_i = \alpha + s_i\beta$ and we wish to test $H_0: \beta = 0$. Let $\mathbf{p} = (p_1 \cdots p_R)^T$ with covariance matrix $\Sigma = \text{diag}(\sigma_1^2 \cdots \sigma_R^2)$ with $\sigma_i^2 = \pi_i(1 - \pi_i)/n_i$. H_0 implies that $\pi_1 = \cdots = \pi_R = \pi$ that is estimated by the marginal proportion p .

The estimates $(\hat{\alpha}, \hat{\beta})$ and their covariance matrix are provided by an application of weighted least squares with a heteroscedastic covariance matrix of errors as described in Section A.5.3. However, since the objective is to develop a test of significance, it is sufficient to estimate β using the estimated covariance matrix evaluated under the null hypothesis $(\hat{\Sigma}_0)$ with the marginal proportion p where $\hat{\Sigma}_0 = p(1 - p) \text{diag}(n_1^{-1} \cdots n_R^{-1})$. The design matrix and response vector then have elements

$$\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ s_1 & \cdots & s_R \end{bmatrix}^T \quad (2.147)$$

$$\mathbf{Y} = [p_1 \cdots p_R]^T.$$

Using (A.70) and (A.71), both with $\hat{\Sigma}_0$, yields estimates of the parameter vector $\boldsymbol{\theta} = (\alpha \ \beta)^T$ and its covariance matrix. These include

$$\hat{\beta} = \frac{\sum_{i=1}^R n_i s_i (p_i - p)}{\sum_{i=1}^R n_i (s_i - \bar{s})^2} \quad (2.148)$$

$$\hat{V}(\hat{\beta}) = \frac{p(1 - p)}{\sum_{i=1}^R n_i (s_i - \bar{s})^2}$$

(see Problem 2.16). The resulting Wald test of H_0 , $X_{CA}^2 = \widehat{\beta}^2 / \widehat{V}(\widehat{\beta})$, simplifies to

$$X_{CA}^2 = \frac{\left[\sum_{i=1}^R n_i s_i (p_i - p) \right]^2}{p(1-p) \sum_{i=1}^R n_i (s_i - \bar{s})^2} = \frac{\left[\sum_{i=1}^R n_i p_i (s_i - \bar{s}) \right]^2}{p(1-p) \sum_{i=1}^R n_i (s_i - \bar{s})^2}, \quad (2.149)$$

where $\bar{s} = \sum_{i=1}^R n_i s_i / N$.

The model specifies that

$$p_i = \widehat{\alpha} + \widehat{\beta} s_i = p + \widehat{\beta} (s_i - \bar{s}) \quad (2.150)$$

and thus

$$\sum_{i=1}^R n_i s_i (p_i - p) = \widehat{\beta} \sum_{i=1}^R n_i s_i (s_i - \bar{s}) = \widehat{\beta} \sum_{i=1}^R n_i (s_i - \bar{s})^2. \quad (2.151)$$

Then the test reduces to

$$X_{CA}^2 = \frac{\widehat{\beta}^2 \sum_{i=1}^R n_i (s_i - \bar{s})^2}{p(1-p)}. \quad (2.152)$$

When used with rank scores ($s_i = r_i$), $\widehat{\beta} = \widehat{\beta}_r$ is the change in p per unit increase in rank. However, the ranks are a function of the sample size. If the fractional ranks are used, say $f_i = r_i / N$, then $\widehat{\beta}_f = \widehat{\beta}_r N$. For a given change in these fractions, say d , the change in p is $d\widehat{\beta}_f$.

This expression can be further simplified with rank scores ($s_i = r_i$) and equally sized groups ($n_i = n = N/R$). The rank score in the i th group then is $r_i = (i-1)n + (n+1)/2$ with mean $\bar{r} = (nR+1)/2$ and deviation

$$r_i - \bar{r} = \frac{n [2i - (R+1)]}{2}. \quad (2.153)$$

Noting that $\sum_{i=1}^R i^2 = R(2R+1)(R+1)/6$, it follows that

$$n \sum_{i=1}^R (r_i - \bar{r})^2 = \frac{n^3 (R+1) R(R-1)}{12}. \quad (2.154)$$

Then

$$X_{CA}^2 = \frac{\widehat{\beta}_r^2 n^3 (R+1) R(R-1)}{12p(1-p)}. \quad (2.155)$$

If the fractional ranks are used then $\widehat{\beta}_f = \widehat{\beta}_r n R$. Stated in terms of $\widehat{\beta}_f$ yields

$$X_{CA}^2 = \frac{\widehat{\beta}_f^2 n (R+1) (R-1)}{12Rp(1-p)}. \quad (2.156)$$

Since the rank score assigned to each group differs by n , and the fractional rank by $1/R$, then the change in p per unit change in group is $\widehat{\beta}_g = \widehat{\beta}_f / R$.

Example 2.21 *Coronary Heart Disease in the Framingham Study (continued)*

We now use the Cornfield (1962) Framingham Study data to determine whether there is a linear trend in the probability of coronary heart disease as the level of cholesterol increases. The data are

CVD	Cholesterol (mg/dL)				Total
	< 200	200–219	220–259	260+	
Yes #	12	8	31	41	92
%	3.8%	3.2%	6.6%	14.3%	6.9%
Total	319	254	470	286	1329

Using the table scores 1 to 4 for the cholesterol categories yields the estimated slope $\hat{\beta} = 0.033092$ and $\hat{V}_{\hat{\beta}} = 4.185 \times 10^{-5}$ and trend test Z -value of -5.51154 or a chi-square test value of $X_{CA}^2 = 26.167$ on 1 df , with $p < 0.001$. Thus, per change in cholesterol group (e.g., 1 to 2), the probability of CVD increases by 0.033.

Using rank scores yields the estimated slope $\hat{\beta}_r = 9.960 \times 10^{-5}$ and $\hat{V}_{\hat{\beta}_r} = 3.561 \times 10^{-10}$ and trend test Z -value of -5.2782 . The corresponding slope in terms of fractional ranks is $\hat{\beta}_f = 0.13237$ and $\hat{V}_{\hat{\beta}_f} = 6.290 \times 10^{-4}$. Then for a 10% difference in rank, $d = 0.10$, the change in p is 0.0133.

For both table and rank scores, the trend Z -test computations are provided by SAS PROC FREQ using the statement

```
proc freq; table died*chol/ trend scores=rank;
```

However, the slope estimates are not provided. These computations are provided by the program *Cornfield trend.sas*.

2.11.7 Exact Tests

The preceding tests of significance for polychotomous and ordinal data are all appropriate for use with large samples. With small samples an exact test is preferred and such tests are provided by StatXact and the SAS PROC FREQ with the *exact* option, among others, to which the reader is referred for examples and computational details.

2.12 PROBLEMS

- 2.1** In Section 2.1.3 the logit was introduced as a way of constructing confidence limits for a proportion that are contained in the unit interval (0,1). Another convenient function, widely used in survival analysis, is the *complementary log-log* transformation $\theta = g(\pi) = \log(-\log \pi)$.

2.1.1. Use the δ -method in conjunction with Slutsky's theorem to show that the asymptotic distribution of $\hat{\theta} = g(p) = \log(-\log p)$ is

$$\log(-\log p) \xrightarrow{d} \mathcal{N} \left[\log(-\log \pi), \frac{1 - \pi}{N\pi (\log \pi)^2} \right], \quad (2.157)$$

where

$$V[\log(-\log p)] \cong \frac{1 - \pi}{N\pi (\log \pi)^2}. \quad (2.158)$$

that can be consistently estimated as

$$\hat{V}[\log(-\log p)] = \frac{1 - p}{Np (\log p)^2}. \quad (2.159)$$

2.1.2. From this result, derive the expression in (2.19) for the asymmetric $(1 - \alpha)$ -level confidence limits for π based on taking the inverse function of the confidence limits for $\theta = \log(-\log \pi)$.

2.1.3. For the case where $x = 2$ and $n = 46$, compute these asymmetric confidence limits and compare them to the values presented in Example 2.1 using the exact computation and those based on the inverse logit transformation.

2.2 Derive the expression for the test-based confidence limits for a single proportion presented in (2.20). Also apply this expression to compute the confidence limits for π for the sample with 2/46 positive outcomes.

2.3 Consider the case of zero events in a sample of size N described in Section 2.4.

2.3.1. Show that the upper one-sided confidence limit equals (2.23).

2.3.2. Show that the sample size required to have an upper confidence limit of π_u at confidence level $1 - \alpha$ is provided by

$$N = \frac{\log(\alpha)}{\log(1 - \pi_u)}. \quad (2.160)$$

2.3.3. For some new drugs, the total number of patients exposed to the agent may be only $N = 500$ patients. With this sample size, if there were no drug-related deaths, show that the 95% confidence interval for the probability of drug-related deaths is $(0, 0.00597)$ or as high as 6 per 1000 patients treated. Given that some drugs are administered to hundreds of thousands of patients each year, is this reassuring?

2.3.4. What is the upper 80% confidence limit?

2.3.5. What sample size would be required to provide an upper 95% confidence limit of 1 in 10,000?

2.4 Show that the cell probabilities (π_1, π_2) in the 2×2 table are a function of the odds ratio and relative risk as follows:

2.4.1. For a given value of π_2 and a given value of the odds ratio OR , show that the corresponding value of π_1 is

$$\pi_1 = \frac{(OR)\pi_2}{(1 - \pi_2) + (OR)\pi_2}. \quad (2.161)$$

2.4.2. Show that the estimated OR and RR are related as follows:

$$\widehat{RR} = \widehat{OR} \left[\frac{1 + (b/d)}{1 + (a/c)} \right]. \quad (2.162)$$

2.4.3. Show that \widehat{OR} approximates \widehat{RR} when either the positive or negative outcome is rare in the population sampled, assuming that m_1/N is small (close to zero).

2.5 Starting from the asymptotic bivariate normal distribution of (p_1, p_2) with covariance matrix as in (2.44):

2.5.1. Use the δ -method to derive the expression for the large sample variance of $\widehat{\theta}$, say $\sigma_{\widehat{\theta}}^2$, and a consistent estimator of this variance, $\widehat{\sigma}_{\widehat{\theta}}^2$, for:

1. The risk difference presented in (2.26).
2. Log relative risk in (2.37).
3. The log odds ratio in (2.46).

2.5.2. For the log relative risk scale show that the estimated variance $\widehat{\sigma}_{\widehat{\theta}}^2$ can be expressed as shown in (2.38).

2.5.3. For the log odds scale, show that the estimated variance $\widehat{\sigma}_{\widehat{\theta}}^2$ is Woolf's estimated variance in (2.47).

2.5.4. In practice, the estimated variance of the relative risk and of the odds ratio (without the log transformation) should not be used for computations of confidence limits. In the following, they are derived principally as an exercise, although similar results are employed in Chapter 9. Apply the δ -method, starting from the results in Problem 2.5.1, to show that the variance of the relative risk (without the log transformation) is asymptotically

$$V(\widehat{RR}) \cong (RR)^2 V[\log(\widehat{RR})] \quad (2.163)$$

and of the odds ratio is

$$V(\widehat{OR}) \cong (OR)^2 V[\log(\widehat{OR})]. \quad (2.164)$$

2.5.5. Under the null hypothesis it is assumed that $H_0: \pi_1 = \pi_2 = \pi$, which is equivalent to $H_0: \theta = \theta_0$, where θ_0 is the null value, $\theta_0 = 0$ for the risk difference, the log relative risk, and log odds ratio. Under the null hypothesis, derive the corresponding expression for the variance of $\widehat{\theta}$ and a consistent estimator thereof for each of these measures.

2.5.6. Redefining the relative risk herein as $RR_{2:1} = \pi_2/\pi_1$, show that

$$V[\log(\widehat{RR}_{2:1})] = V[\log(\widehat{RR}_{1:2})]. \quad (2.165)$$

2.5.7. Also show that the confidence limits for $\log(RR_{2:1})$ are the negatives of those for $\log(RR_{1:2})$, and that the confidence limits for $RR_{2:1}$ are the reciprocals of those for $RR_{1:2}$.

2.6 Starting from the product-binomial likelihood in (2.53), under the null hypothesis $H_0: \pi_1 = \pi_2 = \pi$, show the following:

2.6.1. The probability of the 2×2 table is

$$P(a, b | n_1, n_2, \pi) = \binom{n_1}{a} \binom{n_2}{b} \pi^{m_1} (1 - \pi)^{m_2}. \quad (2.166)$$

2.6.2. In group 1, show that

$$E(a) = \pi n_1 \hat{=} \frac{m_1 n_1}{N} \quad (2.167)$$

and that

$$V(a) = \pi(1 - \pi)n_1 \hat{=} \frac{m_1 m_2 n_1}{N^2}, \quad (2.168)$$

where $\hat{=}$ means "estimated as."

2.6.3. For Cochran's unconditional test, derive the variance $V[a - \hat{E}(a)]$ in (2.91) and its estimate in (2.92).

2.6.4. Use the conditioning arguments in Section 2.4.2 to derive the conditional hypergeometric likelihood in (2.60).

2.6.5. Under the null hypothesis with odds ratio $\varphi = 1$, show that this likelihood reduces to the expression in (2.73).

2.6.6. Using basic probability theory, show that the probability

$$P(a | n_1, m_1, N) = \frac{\binom{n_1}{a} \binom{N - n_1}{m_1 - a}}{\binom{N}{m_1}} \quad (2.169)$$

can be derived as the probability of choosing m_1 of N elements that includes a of n_1 positive elements, for fixed N and n_1 .

2.6.7. Since $\sum_a P(a | n_1, m_1, N) = 1$, show that

$$\sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{N - n_1}{m_1 - i} = \binom{N}{m_1}. \quad (2.170)$$

2.6.8. Then derive the simplifications in (2.74).

2.6.9. Using these results, show that $E(a) = m_1 n_1 / N$ as presented in (2.87).

Hint: Note that for total sample size $N - 1$, the sum of the possible hypergeometric probabilities equals 1.

2.6.10. Show that

$$E[a(a - 1)] = \frac{m_1(m_1 - 1)n_1(n_1 - 1)}{N(N - 1)}. \quad (2.171)$$

Given that $E(a^2) = E[a(a - 1)] + E(a)$, show that the conditional hypergeometric variance is $V_c(a)$ in (2.88).

2.7 Given a sample of N observations with fixed margins (n_1, n_2, m_1, m_2) :

2.7.1. Show that the expected value of the odds ratio, say $\tilde{\varphi}$, can be defined in terms of the probability of the outcome in group 1, π_1 , as

$$\tilde{\varphi} = \frac{\pi_1(n_2 - m_1 + \pi_1 n_1)}{(1 - \pi_1)(m_1 - \pi_1 n_1)}, \quad (2.172)$$

where $E(a) = n_1 \pi_1$, $E(b) = m_1 - n_1 \pi_1$, and so on.

2.7.2. Show that the quadratic equation in π_1 as a function of $\tilde{\varphi}$ is

$$[n_1(\tilde{\varphi} - 1)] \pi_1^2 - [\tilde{\varphi}(n_1 + m_1) + (n_2 - m_1)] \pi_1 + \tilde{\varphi} m_1 = 0. \quad (2.173)$$

2.7.3. Then show that the value of π_1 which satisfies this equation equals

$$\pi_1 = \frac{C}{2[n_1(\tilde{\varphi} - 1)]} \quad (2.174)$$

with

$$C = \varphi(n_1 + m_1) + (n_2 - m_1) - \sqrt{[\tilde{\varphi}(n_1 + m_1) + (n_2 - m_1)]^2 - 4\tilde{\varphi}m_1[n_1(\tilde{\varphi} - 1)]} \quad (2.175)$$

(Thomas and Gart, 1977).

2.7.4. Also show that given the fixed margins, the corresponding value of π_2 that satisfies this relationship is

$$\pi_2 = \frac{m_1 - n_1 \pi_1}{n_2}. \quad (2.176)$$

Thus, one can determine the probabilities $\hat{\pi}_{1u}$ and $\hat{\pi}_{1\ell}$ corresponding to the exact confidence limits on the odds ratio, from which the values of $\hat{\pi}_{2u}$ and $\hat{\pi}_{2\ell}$ can be obtained. These then provide exact limits on the risk difference and relative risk, such as $\widehat{RR}_L = \hat{\pi}_{1\ell}/\hat{\pi}_{2\ell}$.

2.7.5. For the data in Example 2.7, given the exact 95% confidence limits for the odds ratio of (0.0127, 1.0952), use the above expressions to determine the corresponding probabilities. Then show that the exact confidence limits for the risk difference are (-0.6885, 0.0227) and for the relative risk are (0.2890, 1.0452).

2.8 Consider the large sample tests for the 2×2 table.

2.8.1. Show that the Pearson X_P^2 in (2.84) equals Cochran's test X_u^2 in (2.93).

2.8.2. Also show that X_u^2 equals Z^2 , where Z is the usual Z -test for two proportions in (2.83).

2.9 Now consider a test based on some smooth function $g(p)$, such as a test for the log relative risk or log odds ratio with H_0 : $g(\pi_1) = g(\pi_2)$ or, equivalently, H_0 :

$\pi_1 = \pi_2$. The corresponding Z -test then is of the form

$$Z_g = \frac{g(p_1) - g(p_2)}{\sqrt{\widehat{V}[g(p_1) - g(p_2) | H_0]}}. \quad (2.177)$$

Using Taylor's expansion, show that Z_g is asymptotically equal to the usual Z -test in (2.80). *Hint:* Evaluate $g(p_j)$ about the assumed common π under H_0 .

2.10 Consider the population attributable risk (*PAR*) in (2.111) and the consistent estimate in (2.112). Note that the *PAR* is a proportion.

2.10.1. For α_1 specified (fixed), show that the *PAR* estimated odds are

$$\frac{\widehat{PAR}}{1 - \widehat{PAR}} = \frac{\alpha_1(p_1 - p_2)}{p_2}. \quad (2.178)$$

2.10.2. Using the δ -method, show that $V\left(\log\left[\widehat{PAR}/(1 - \widehat{PAR})\right]\right)$ is as expressed in (2.113) and that this variance can be estimated consistently as shown in (2.114).

2.11 Hypothetically, in a clinical trial of ursodiol (a bile acid drug) versus a placebo for the treatment of patients with gallstones, the following 2×2 table was obtained, giving the frequencies of a positive response (disappearance of gallstones) versus no response after 12 months of treatment:

		Group		
		Urso	Placebo	
Response	+	23	8	
	-	87	92	
		110	100	210

(2.179)

2.11.1. For this table, compute the estimates of the risk difference, relative risk and its logarithm, and the odds ratio and its logarithm. Also compute the estimated variance and standard error, and the 95% confidence intervals for the risk difference, log relative risk, and the log odds ratio.

2.11.2. From the log odds ratio and log relative risk, also compute the asymmetric 95% confidence limits on the odds ratio and relative risk, respectively.

2.11.3. Also compute the asymmetric confidence limits for the relative risk and odds ratio for the placebo group to the ursodiol group; e.g., $RR_{2:1}$.

2.11.4. Then compute the Pearson contingency chi-square test (Cochran's test) and the Mantel-Haenszel test, with and without the continuity correction.

2.11.5. Now use the results from the Mantel-Haenszel conditional test without the continuity correction to compute test-based confidence limits on the log odds ratio and the log relative risk. From these compute the asymmetric confidence limits on the odds ratio and the relative risk. Compare these to the asymmetric confidence limits obtained above.

2.11.6. Now use PROC FREQ in SAS with the *cmh chisq* options to perform an analysis of the above 2×2 table. This will require that you input the data as shown in Table 2.2. Alternatively, use a program such as StatXact to compute the exact *p*-values. Compare the respective large sample tests to the exact two-sided *p*-value.

2.11.7. Compute the number needed to treat with ursodiol (*NNT*) so as to yield one treatment success. Use the 95% confidence limits on the risk difference to compute the confidence limits on the *NNT*.

2.11.8. Now suppose that the control group received the current standard of care rather than placebo and that it was desired to demonstrate noninferiority one-sided at the 95% level of confidence. Using a relative risk with a margin of 10%, do these data demonstrate noninferiority?

2.12 Mogensen (1984) examined the influence of early diabetic nephropathy on the premature mortality associated with type 2 or adult onset diabetes mellitus in a sample of 204 men 50 to 75 years of age with onset of diabetes after 45 years of age. At the time of entry into the cohort, each subject was characterized as having normal levels of albumin excretion defined as AER $<30 \mu\text{g}/\text{min}$ versus those with microalbuminuria ($30 \leq \text{AER} < 140 \mu\text{g}/\text{min}$). See Section 1.5 for a description of these terms. The cohort was then followed for 10 years. The following results were obtained:

		Albuminuria Group			
		Micro	Normal		
Died	59	55	114		(2.180)
	17	73	90		
		76	128	204	

(reproduced with permission).

2.12.1. Use the relative risk (risk ratio) for mortality, with an appropriate confidence interval, and the Mantel-Haenszel test to describe the relationship between the presence of microalbuminuria versus normal albumin levels and the 10-year mortality.

2.12.2. Compute the population attributable risk and its 95% confidence limits. Describe these results in terms of the impact of diabetic nephropathy on 10 year mortality among those with type 2 diabetes mellitus.

2.13 Starting from the unconditional product-binomial likelihood (2.53) we now use a *logit (logistic) model* to derive the conditional hypergeometric distribution for a 2×2 table. Conditional on n_1 and n_2 (or n_1, N), under the logistic model we can express the logit of π_1 and π_2 as

$$\log \left(\frac{\pi_1}{1 - \pi_1} \right) = \alpha + \beta, \quad \log \left(\frac{\pi_2}{1 - \pi_2} \right) = \alpha. \quad (2.181)$$

2.13.1. Show that the *inverse logits* are provided by logistic functions of the form

$$\pi_1 = \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}}, \quad \pi_2 = \frac{e^\alpha}{1 + e^\alpha} \quad (2.182)$$

and

$$1 - \pi_1 = \frac{1}{1 + e^{\alpha+\beta}}, \quad 1 - \pi_2 = \frac{1}{1 + e^\alpha}. \quad (2.183)$$

2.13.2. Then show that

$$\varphi = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = e^\beta. \quad (2.184)$$

2.13.3. Substituting this logistic model representation into (2.53), show that the probability of the 2×2 table can be expressed as

$$P(a, m_1 | n_1, n_2) = \binom{n_1}{a} \binom{n_2}{m_1 - a} \frac{e^{a\beta} e^{m_1 \alpha}}{(1 + e^{\alpha+\beta})^{n_1} (1 + e^\alpha)^{n_2}}, \quad (2.185)$$

where a and m_1 are random.

2.13.4. Conditioning on m_1 , show that

$$\begin{aligned} P(a | m_1) &= \frac{P(a, m_1)}{\sum_{i=a_\ell}^{a_u} P(i, m_1)} = \frac{P(a, m_1)}{P(m_1)} \\ &= \frac{\binom{n_1}{a} \binom{n_2}{m_1 - a} e^{a\beta}}{\sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{n_2}{m_1 - i} e^{i\beta}}, \end{aligned} \quad (2.186)$$

where (a_ℓ, a_u) are the limits for the possible values for the index frequency a .

2.13.5. Show that this expression is equivalent to (2.60).

2.14 Consider a set of Bernoulli variables (x_{i1}, \dots, x_{iC}) that denotes whether the i th element in a sample of N observations falls in category $1, \dots, C$ of some characteristic.

2.14.1. Using basic principles, derive the covariance matrix of the vector $(x_{i1} \dots x_{iC})$ as shown in Example A.2.

2.14.2. Show by induction that the inverse of the covariance matrix for the vector of $C - 1$ frequencies under a multinomial satisfies the expression in (2.123). Consider the case of $C = 4$.

2.14.3. Show that the large sample Wald test for a multinomial with specified probabilities in each category equals the Pearson chi-square test as stated in (2.124).

2.14.4. Similar to the derivations in Example A.4, derive the expressions for the variances and covariances of the continuation-ratio logits presented in (2.129).

2.15 Consider the case of a $2 \times C$ contingency table formed by two groups with C proportions in each.

2.15.1. Show that the T^2 -like test in (2.134) equals the Pearson test in (2.135).

2.15.2. Derive the variances and covariances of the generalized odds ratios as shown in (2.136).

2.16 Consider the case of R groups formed from an ordinal or quantitative variable with a binary response that is assumed to be a linear function of the score assigned to each group.

2.16.1. Using weighted least squares with the covariance matrix evaluated under H_0 , derive the expressions for the slope $\hat{\beta}$ and its estimated variance shown in (2.148), and the Cochran-Armitage test of trend presented in (2.149).

2.16.2. Derive the simplification of the test presented in (2.155) for the case of equal-sized groups with rank scores.

2.17 Diabetic retinopathy (DR) was graded for severity on an ordinal scale with categories 1 = none, 2 = very mild nonproliferative DR (NPDR), 3 = mild NPDR, 4 = moderate NPDR, 5 = severe NPDR, 6 = mild proliferative DR (PDR), 7 = moderate PDR, 8 = high risk (sight threatening) PDR. The following table shows the numbers of subjects at each level of retinopathy among those evaluated at five years of follow-up in the DCCT intensive versus conventional therapy groups.

Group	Retinopathy							
	1	2	3	4	5	6	7	8
Intensive	121	246	113	37	8	2	8	2
Conventional	79	214	142	59	14	8	9	18

2.17.1. Compute a rank score test of the differences between the two groups.

2.17.2. Compute the continuation-ratio log odds ratios and the variance (standard error) of each.

Sample Size, Power, and Efficiency

Chapter 2 described the large sample distribution of estimates of the differential in risk between two groups, and tests of the significance of these differences. In this chapter we consider the evaluation of the sample size required for a study based on either the desired precision of an estimate, such as of the risk difference, or based on the power function of a test, such as the Z -test for two proportions. We begin with the derivation of the power function for the general family of Z -tests, which will also be employed in subsequent chapters to evaluate the power of other normal deviate Z -tests. The developments are expanded to include the family of chi-square tests, such as the Pearson contingency test for an $R \times C$ table. This is followed by an introduction of the concepts of asymptotic efficiency under local alternatives that is used in subsequent chapters to derive asymptotically efficient, or most powerful, tests.

There is a considerable literature on methods for evaluation of the precision of estimates of parameters and the power of specific statistical tests. Since estimation precision and the power of a test largely depend on sample size, consideration of these properties allows the determination of an adequate, if not optimal, sample size when a study is designed. Interested readers are referred to the following references for additional materials. McHugh and Le (1984) present a review of procedures for determination of sample size from the perspective of the precision of an estimator; while Lachin (1981, 1998), and Donner (1984), among others, review methods for sample size determination from the perspective of the power of commonly used statistical tests. General reference texts on the topic include Machin and Campbell (1987), Cohen (1988), Desu and Raghavarao (1990), Schuster (1990), and Odeh and Fox (1991), among others.

3.1 ESTIMATION PRECISION

Consider that we wish to estimate a parameter θ in a population, such as the log odds ratio, based on a simple random sample that yields an estimator $\hat{\theta}$ that is normally distributed, at least asymptotically, as $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$. In most instances, the large sample variance σ^2 can be expressed as a function of the sample size N , such as $\sigma^2 = \phi^2/N$, where ϕ^2 is a *variance component*. Then the $(1 - \alpha)$ -level confidence interval (C.I.) for θ is of the form $\hat{\theta} \pm e_\alpha$, where the precision of the estimate at level $1 - \alpha$ is

$$e_\alpha = Z_{1-\alpha/2}\sigma \quad (3.1)$$

and $Z_{1-\alpha/2}$ is the upper two-sided standard normal deviate at level $1 - \alpha/2$. Since σ is a function of the sample size N , then so also is the $(1 - \alpha)$ -level precision of the estimate e_α . The smaller the value of e_α , the greater the precision of the estimate.

This relationship also can be inverted. For any stated degree of precision e of the estimate, the confidence level $1 - \alpha$ is provided by the standardized deviate $Z_{1-\alpha/2} = e/\sigma = \sqrt{N}e/\phi$. This allows one to evaluate the relationship between the level of confidence and the degree of precision of the estimate for different sample sizes. Alternatively, if the sample size is under the control of the investigator when a study is designed or planned, the above relationship can be inverted to yield the sample size N required to provide a confidence interval estimate with desired precision $\pm e$ at level $1 - \alpha$ for a given variance component ϕ . This is given by

$$N = \left(\frac{Z_{1-\alpha/2}\phi}{e} \right)^2. \quad (3.2)$$

Likewise, one can consider the estimation precision and sample size for a study comparing two populations (groups) with respect to the risk difference, the log relative risk, or the log odds ratio, using the large sample variance and the asymptotic normal distribution for each derived in Chapter 2. Those interested in other common instances are referred to the article by McHugh and Le (1984) or to texts on sampling.

Such computations assume that the variance is known or specified a priori. In the case of functions of probabilities, such as the log odds ratio, this requires that the probabilities in each population are known or specified. This approach can be further generalized to allow for sampling variation in the estimated variance, and thus in the precision of the estimate, by determining the sample size needed to provide probability $1 - \beta$ that the realized $1 - \alpha$ confidence interval will have precision no greater than e (cf. Kupper and Hafner, 1989).

Example 3.1 Simple Proportion

For example, the simple proportion with a characteristic of interest, say p , from a sample of N observations is asymptotically distributed as $p \xrightarrow{d} \mathcal{N}[\pi, \pi(1 - \pi)/N]$, where π is the probability of the characteristic in the population and the large sample variance of p is $\sigma^2 = \pi(1 - \pi)/N$ with $\phi^2 = \pi(1 - \pi)$. To estimate a probability π that is assumed to be ≤ 0.3 (or $\pi \geq 0.7$) with precision $e = 0.02$ at 90%

confidence, the required sample size is $N = (1.645/0.02)^2(0.3 \times 0.7) = 1421$ (rounded up to the next whole integer). If this sample size is too large but a sample size of 1000 is feasible, then one can determine that a 90% confidence interval with $N = 1000$ provides a degree of precision of $e = 1.645\sqrt{(0.3 \times 0.7)/1000} = 0.024$. Alternatively, a sample size of $N = 1000$ provides 83% confidence of estimating π with a precision of $e = 0.02$, where the corresponding normal deviate is $Z_{1-\alpha/2} = \sqrt{1000}(0.02)/\sqrt{0.3 \times 0.7} = 1.38$.

3.2 POWER OF Z-TESTS

Now consider an investigation that is designed to determine which of two populations is superior, or whether one population is superior to the other. To reach an inference as to the difference between the two populations, we conduct a statistical test of significance and also present the large sample confidence interval. The test statistic assesses the probability that the observed results could have occurred by chance, whereas the confidence limits provide a description of the precision of the sample estimate of the group difference. In this setting, it is usually important to determine the sample size so as to provide a suitably high probability of obtaining a statistically significant result when a true difference of some magnitude exists between the populations being compared. This probability is termed the *power* of the test. We now derive the general expressions for the power of any normal deviate *Z*-test that can be applied to specific tests, such as the test for two proportions, and the corresponding equations for the sample size needed to achieve a desired level of power.

3.2.1 Type I and II Errors and Power

Consider a *Z*-test based on a statistic T that is a consistent estimator of the mean μ of a normal distribution. Using T , we wish to test the null hypothesis

$$H_0: \mu = \mu_0, \quad (3.3)$$

where usually $\mu_0 = 0$. For now consider an upper-tailed one-sided test of H_0 versus the one-sided alternative hypothesis

$$H_1: \mu = \mu_1 > \mu_0. \quad (3.4)$$

This implies that $H_0: \Delta = \mu_1 - \mu_0 = 0$ versus $H_1: \Delta > 0$. To allow for cases where the variance of T is a function of μ , $V(T)$ is expressed as

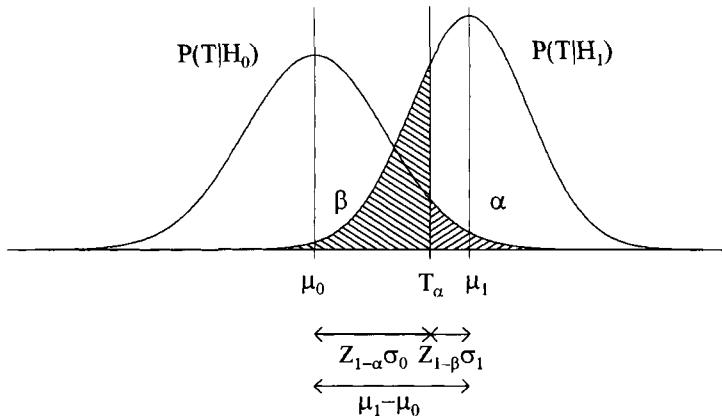
$$V(T) = \sigma^2(\mu). \quad (3.5)$$

We then have two possible distributions of T , that under the null hypothesis H_0 and that under the alternative hypothesis H_1 , designated as:

$$\text{Null: } T_{(H_0)} \sim \mathcal{N}(\mu_0, \sigma_0^2), \quad \sigma_0^2 = \sigma^2(\mu_0) \quad (3.6)$$

$$\text{Alternative: } T_{(H_1)} \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad \sigma_1^2 = \sigma^2(\mu_1), \quad (3.7)$$

Fig. 3.1 Distribution of a test statistic under the null and alternative hypotheses, with the rejection region of size α and a type II error probability of size β .



where $\mu_1 > \mu_0$. An example is depicted in Figure 3.1.

To conduct the test of significance, H_0 is rejected when the observed value of the statistic t falls within the upper-tail rejection region of size α , or

$$t \geq T_\alpha = \mu_0 + Z_{1-\alpha}\sigma_0. \quad (3.8)$$

When the null hypothesis is true, the probability of rejection is only α and the probability of failing to reject it is $1 - \alpha$. Both quantities are determined by the investigator through specification of the size α of the rejection region. However, when the alternative hypothesis H_1 is true, the probabilities of rejecting or failing to reject H_0 are not fixed by the investigator uniquely, but rather, are determined by other factors, including the magnitude of the true difference Δ .

Thus, for any fixed $\Delta = \mu_1 - \mu_0$, two types of errors can occur, a false positive Type I error with probability α and a false negative Type II error with probability β , as presented by the following:

	H_0	$H_1: \mu_1 - \mu_0 = \Delta$
Reject: +	α	$1 - \beta(\Delta, N, \alpha)$
Fail to Reject: -	$1 - \alpha$	$\beta(\Delta, N, \alpha)$
	1.0	1.0

(3.9)

Because α is fixed arbitrarily, it does not depend on either the value of Δ or the value of N . Conversely, the probability of a false negative error, designated as $\beta(\Delta, N, \alpha)$, depends explicitly on Δ and on the total sample size N , as well as the size of the test α . The complement of the Type II error, $1 - \beta(\Delta, N, \alpha)$, is the *power* of the test to detect a difference Δ with total sample size N .

The relationship between α and β is illustrated in Figure 3.1. It is clear that one cannot fix the levels of both the type I and type II error probabilities α and β . Rather, α and β are inversely related. As the critical value T_α is shifted toward μ_0 , α increases and β decreases, and vice versa. Also, as the magnitude of the difference under the alternative Δ increases, the distribution under the alternative is shifted to the right and β decreases, even though α remains a constant. Finally, for fixed Δ , as the variance of the statistic σ^2 decreases, the curves shrink. This is readily accomplished by increasing the sample size. This has two effects. First, in order to retain the upper tail area of size α under the null hypothesis distribution, the value T_α shrinks toward the mean μ_0 as the variance $\sigma^2(\mu_0)$ becomes smaller. Second, as the variance $\sigma^2(\mu_1)$ decreases under the alternative hypothesis, the curve under H_1 also shrinks. Each factor contributes to a decrease in the value of β . Thus, while the value of the Type I error probability is specified a priori, the value for the Type II error probability $\beta(\Delta, N, \alpha)$ is determined by other factors.

We now describe these relationships algebraically. Under the null hypothesis, the significance test leads to rejection of H_0 if the standardized Z -test satisfies

$$z = \frac{t - \mu_0}{\sigma_0} \geq Z_{1-\alpha}. \quad (3.10)$$

Therefore, the Type II error probability is

$$\beta = P[z < Z_{1-\alpha} \mid H_1] \quad (3.11)$$

and its complement, power, is

$$1 - \beta = P[z \geq Z_{1-\alpha} \mid H_1], \quad (3.12)$$

where each is evaluated under the alternative hypothesis distribution. These quantities are identical to the areas depicted in Figure 3.1 with a change of scale from T to Z .

Under $T_{(H_1)}$ in (3.7), where $\Delta = \mu_1 - \mu_0 > 0$, then

$$T - \mu_0 \sim \mathcal{N}[\mu_1 - \mu_0, \sigma_1^2] \quad (3.13)$$

and

$$Z = \frac{T - \mu_0}{\sigma_0} \sim \mathcal{N}\left[\frac{\mu_1 - \mu_0}{\sigma_0}, \frac{\sigma_1^2}{\sigma_0^2}\right]. \quad (3.14)$$

Therefore,

$$\beta = P[z < Z_{1-\alpha} \mid H_1] = \Phi\left[\frac{Z_{1-\alpha} - \left(\frac{\mu_1 - \mu_0}{\sigma_0}\right)}{\sigma_1/\sigma_0}\right] = \Phi[Z_\beta]. \quad (3.15)$$

Thus,

$$Z_\beta = \frac{Z_{1-\alpha} - \left(\frac{\mu_1 - \mu_0}{\sigma_0} \right)}{\sigma_1/\sigma_0}. \quad (3.16)$$

However, β is the area to the left of T_α in Figure 3.1, and we desire the expression for $1 - \beta$ that is the area to the right. Since $Z_{1-\beta} = -Z_\beta$, then

$$\begin{aligned} Z_{1-\beta} &= \left(\frac{\sigma_0}{\sigma_1} \right) \left[\frac{\mu_1 - \mu_0}{\sigma_0} - Z_{1-\alpha} \right] \\ &= \frac{(\mu_1 - \mu_0) - Z_{1-\alpha}\sigma_0}{\sigma_1} \\ &= \frac{\Delta - Z_{1-\alpha}\sigma_0}{\sigma_1}. \end{aligned} \quad (3.17)$$

Thus, $(Z_{1-\beta} > 0) \Rightarrow (1 - \beta > 0.5)$.

For a one-sided left-tailed alternative hypothesis, $H_1: \mu_1 < \mu_0$ with $\mu_1 - \mu_0 = \Delta < 0$, a similar derivation again leads to this result but with the terms $-(\mu_1 - \mu_0)$ and $-\Delta$ (see Problem 3.1). In general, therefore, for a one-sided test in either tail, the basic equation relating Δ , α , and β is of the form

$$|\Delta| = Z_{1-\alpha}\sigma_0 + Z_{1-\beta}\sigma_1 \quad (3.18)$$

for a Δ in either direction under the specified one-sided alternative hypothesis.

This expression can also be derived heuristically, as shown in Figure 3.1. The line segment distance $\Delta = \mu_1 - \mu_0$ is partitioned into the sum of the line segment distance $|T_\alpha - \mu_0| = Z_{1-\alpha}\sigma_0$ under the null distribution given H_0 plus the distance $|T_\alpha - \mu_1| = Z_{1-\beta}\sigma_1$ under the alternative distribution given H_1 .

For a two-sided test, Figure 3.1 would be modified as follows. The distribution under the null hypothesis would have a two-sided rejection region with an area of size $\alpha/2$ in each tail. Then, for a fixed alternative, in this case a positive value for Δ , there would also be a contribution to the type II probability β from the corresponding rejection area in the far left tail under the alternative hypothesis distribution H_1 . For an alternative such as that shown in the figure, where μ_1 is a modest distance from μ_0 , this additional contribution to the probability β is negligible and can be ignored. In this case, the general expression (3.18) is obtained with the value $Z_{1-\alpha/2}$.

3.2.2 Power and Sample Size

The general expression (3.18) provides an explicit equation relating Δ , N , and $1 - \beta$ that forms the basis for the evaluation of power and the determination of sample size for any test statistic that is asymptotically normally distributed. If we can decompose the variance expressions such that $\sigma_0^2 = \phi_0^2/N$ and $\sigma_1^2 = \phi_1^2/N$, we obtain the set of

equations

$$|\Delta| = \frac{Z_{1-\alpha}\phi_0 + Z_{1-\beta}\phi_1}{\sqrt{N}} \quad (3.19)$$

$$Z_{1-\beta} = \frac{\sqrt{N}|\Delta| - Z_{1-\alpha}\phi_0}{\phi_1} \quad (3.20)$$

$$N = \left[\frac{Z_{1-\alpha}\phi_0 + Z_{1-\beta}\phi_1}{\Delta} \right]^2. \quad (3.21)$$

Equation (3.20) is used to compute the power function for given N and Δ , while (3.21) allows the a priori determination of the N that provides power $1 - \beta$ to detect a difference Δ with an α -level test. For a two-sided test, $Z_{1-\alpha/2}$ is employed.

Lachin (1981) presented a simplification of these general expressions for cases where the variances are approximately equal, that is, $\sigma_0^2 \doteq \sigma_1^2 \doteq \sigma^2 = \phi^2/N$. Substituting into (3.18) yields the simplified general expressions

$$|\Delta| = (Z_{1-\alpha} + Z_{1-\beta}) \frac{\phi}{\sqrt{N}} \quad (3.22)$$

$$\sqrt{N}\tau = Z_{1-\alpha} + Z_{1-\beta} \quad (3.23)$$

as a function of the *noncentrality factor*

$$\tau = \frac{|\Delta|}{\phi}. \quad (3.24)$$

This in turn yields simplifications of the equations for power and sample size

$$Z_{1-\beta} = \tau\sqrt{N} - Z_{1-\alpha} \quad (3.25)$$

$$N = \left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{\tau} \right)^2. \quad (3.26)$$

Lachin (1981) used these to derive equations for the power of many commonly used tests, such as the test for mean values, proportions, paired mean differences, paired proportions, exponential hazard ratios, and correlations, among others.

As will be shown in Section 3.4, these expressions also arise from a consideration of the *noncentrality parameter* that is the expected value of the Z statistic under the alternative hypothesis, or $\psi = E(Z | H_1)$. When the variances are approximately equal, from (3.14),

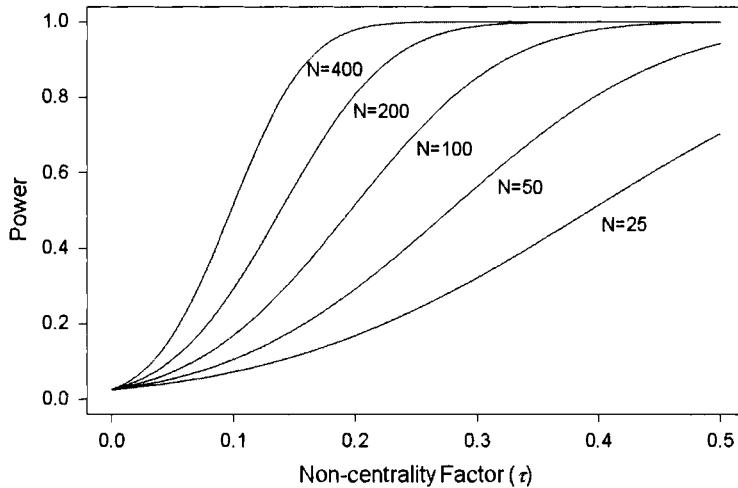
$$|\psi| = \frac{|\mu_1 - \mu_0|}{\sigma} \doteq \frac{|\mu_1 - \mu_0|}{\phi/\sqrt{N}} = \sqrt{N}\tau, \quad (3.27)$$

with noncentral factor τ . Then the basic equation (3.18) is based on the relationship

$$|\psi| = Z_{1-\alpha} + Z_{1-\beta}, \quad (3.28)$$

which relates the noncentrality parameter of the test to the levels of α and β .

Fig. 3.2 Power as a function of the noncentrality factor (τ) and the total sample size (N) for $\alpha = 0.05$, two-sided.



In general, these relationships can be expressed using a set of power curves as in Figure 3.2. Such curves depict the increase in power as the value of Δ (or τ) increases, and also as N increases. For a two-sided test, the power curves are symmetric about $\Delta = 0$. Such power functions are computed using (3.20) over a range of values for Δ and N , given the type I error probability α and values for the other parameters involved, such as the variance components ϕ_0 and ϕ_1 . This figure presents the power function for a one-sided test at level $\alpha = 0.05$ of the differences between two means with unit variance ($\phi^2 = 1$). Equation (3.21), therefore, simply determines the sample size N that provides a power function (curve) that passes through the point $(\Delta, 1 - \beta)$. The essential consideration is that there is a power function associated with any given sample size N that describes the level of power $1 - \beta$ with which any difference Δ can be detected.

3.3 TEST FOR TWO PROPORTIONS

The Z -test for two proportions described in Section 2.6.1 falls in the class of tests employed in the preceding section. Thus, the above general expressions can be used to derive the specific expressions required to evaluate the power of this test and to determine the required sample size. Because the square of the Z -test is algebraically equal to the contingency chi-square test (see Problem 2.8), the power of the latter is also provided by the power of the two-sided Z -test.

3.3.1 Power of the Z-Test

The Z -test for two proportions in (2.83) is based on the test statistic $T = p_1 - p_2$, where $E(T) = \pi_1 - \pi_2$. Under $H_1: \pi_1 \neq \pi_2$, from (3.7), $T \stackrel{d}{\approx} \mathcal{N}[\mu_1, \sigma_1^2]$ with $\mu_1 = \pi_1 - \pi_2$ and

$$\sigma_1^2 = \frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}. \quad (3.29)$$

Under $H_0: \pi_1 = \pi_2 = \pi$, then from (3.6), $T \stackrel{d}{\approx} \mathcal{N}[\mu_0, \sigma_0^2]$ with $\mu_0 = 0$ and

$$\sigma_0^2 = \pi(1 - \pi) \left(\frac{1}{n_1} + \frac{1}{n_2} \right). \quad (3.30)$$

To allow for unequal sample sizes, denote the expected sample fraction in the i th group as ξ_i , where $\xi_i = E(n_i/N)$ ($i = 1, 2$) and $\xi_1 + \xi_2 = 1$. Then, the variances can be factored as

$$\sigma_0^2 = \frac{\phi_0^2}{N}, \quad \phi_0^2 = \pi(1 - \pi) \left(\frac{1}{\xi_1} + \frac{1}{\xi_2} \right) \quad (3.31)$$

and

$$\sigma_1^2 = \frac{\phi_1^2}{N}, \quad \phi_1^2 = \frac{\pi_1(1 - \pi_1)}{\xi_1} + \frac{\pi_2(1 - \pi_2)}{\xi_2}. \quad (3.32)$$

Therefore, $\Delta = \mu_1 - \mu_0 = \pi_1 - \pi_2$, and from (3.18) the basic equation relating the size of the test α , the power $1 - \beta$, and sample size N is

$$\sqrt{N} |\mu_1 - \mu_0| = Z_{1-\alpha} \sqrt{\pi(1 - \pi) \left(\frac{1}{\xi_1} + \frac{1}{\xi_2} \right)} \quad (3.33)$$

$$+ Z_{1-\beta} \sqrt{\frac{\pi_1(1 - \pi_1)}{\xi_1} + \frac{\pi_2(1 - \pi_2)}{\xi_2}} \\ = Z_{1-\alpha}\phi_0 + Z_{1-\beta}\phi_1, \quad (3.34)$$

where $Z_{1-\alpha/2}$ is employed for a two-sided test. Since the Z -test in (2.83) employs the variance estimated under the null hypothesis, where $p = \xi_1 p_1 + \xi_2 p_2$, then for the evaluation of the above equations, σ_0^2 and ϕ_0^2 are computed using

$$\pi = \xi_1 \pi_1 + \xi_2 \pi_2. \quad (3.35)$$

Thus, only the values π_1 and π_2 need be specified.

Solving for $Z_{1-\beta}$, the level of power provided by a given sample size N to detect the difference between proportions with specified probabilities π_1 and π_2 is obtained from

$$Z_{1-\beta} = \frac{\sqrt{N} |\pi_1 - \pi_2| - Z_{1-\alpha}\phi_0}{\phi_1}. \quad (3.36)$$

Likewise, the sample size required to provide power $1 - \beta$ to detect a difference $\Delta = \pi_1 - \pi_2$ is provided by

$$N = \left[\frac{Z_{1-\alpha}\phi_0 + Z_{1-\beta}\phi_1}{\pi_1 - \pi_2} \right]^2. \quad (3.37)$$

With equal sample sizes $\xi_1 = \xi_2 = 1/2$, these expressions simplify slightly using

$$\begin{aligned}\phi_0^2 &= 4\pi(1 - \pi) \\ \phi_1^2 &= 2\pi_1(1 - \pi_1) + 2\pi_2(1 - \pi_2).\end{aligned}\quad (3.38)$$

Finally, for this test, Lachin (1981) shows that $\sigma_0^2 > \sigma_1^2$, in which case we can use the conservative simplifications presented in (3.22 – 3.24) that yield the noncentrality factor

$$\tau = \frac{|\pi_1 - \pi_2|}{\sqrt{\pi(1 - \pi)(1/\xi_1 + 1/\xi_2)}} = \frac{|\pi_1 - \pi_2|}{\phi_0}. \quad (3.39)$$

While (3.36) and (3.37) are preferred for the assessment of power or sample size, the noncentral factor (3.39) will be used below to explore other relationships.

Example 3.2 Planning a Study for Superiority

For example, suppose that we wish to plan a study with two equal-sized groups ($n_1 = n_2$) to detect a 30% reduction in mortality associated with congestive heart failure with a new drug versus placebo, where the one-year mortality in the control group is assumed to be no greater than 0.40. Thus, $\pi_2 = 0.40$ and $\pi_1 = 0.28$ ($= 0.70 \times 0.40$). Under the null hypothesis we assume that $\pi_1 = \pi_2 = \pi = 0.34$. We desire 90% power for a two-sided test for two proportions at $\alpha = 0.05$. Using (3.37) the required total N is obtained as

$$N = \left[\frac{1.96[4(0.34 \times 0.66)]^{\frac{1}{2}} + 1.282[2(0.28 \times 0.72) + 2(0.4 \times 0.6)]^{\frac{1}{2}}}{0.40 - 0.28} \right]^2 = 652$$

rounded up to the nearest even integer.

Alternatively, one could solve for $Z_{1-\beta}$ to determine the power to detect a difference with a specified sample size, or the magnitude of the difference that could be detected with a given power for a specific sample size. For example, the power to detect this same difference with a smaller sample size of $N = 500$ using (3.36) is provided by

$$Z_{1-\beta} = \frac{\sqrt{500}(0.40 - 0.28) - 1.96[4(0.34 \times 0.66)]^{\frac{1}{2}}}{[2(0.28 \times 0.72) + 2(0.40 \times 0.60)]^{\frac{1}{2}}} = 0.879,$$

yielding 81% power.

Note that for a fixed N and a fixed value of π_2 , such computations could be used to generate the power curve as a function of increasing values of π_1 (or Δ). However, as π_1 changes, so do the values of π from (3.35).

Table 3.1 Noncentrality factors for a test of the difference between two proportions with probabilities π_1 and π_2 .

$\pi_2 \setminus \pi_1$	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.140	0.250	0.346	0.436	0.524	0.612	0.704	0.800
0.2	—	0.115	0.218	0.314	0.408	0.503	0.600	0.704
0.3	—	—	0.105	0.204	0.302	0.40	0.503	0.612
0.4	—	—	0.101	0.20	0.302	0.408	0.524	—
0.5	—	—	—	0.101	0.204	0.314	0.436	—
0.6	—	—	—	—	0.105	0.218	0.346	—
0.7	—	—	—	—	—	0.115	0.250	—
0.8	—	—	—	—	—	—	0.140	—

3.3.2 Relative Risk and Odds Ratio

Since the test statistic is $T = p_1 - p_2$, then the power function is naturally expressed in terms of the risk difference. However, in many cases, it is more desirable to express power as a function of the odds ratio or relative risk. To do so, one specifies the probability in the control group π_2 and then determines the value of π_1 in the exposed or treated group which corresponds to the specified relative risk or odds ratio.

Since the relative risk is $RR = \pi_1/\pi_2$, then for any value of π_2

$$\pi_1 = \pi_2(RR). \quad (3.40)$$

Likewise, since the odds ratio is $OR = \pi_1(1 - \pi_2)/[\pi_2(1 - \pi_1)]$, then solving for π_1 yields

$$\pi_1 = \frac{(OR)\pi_2}{(1 - \pi_2) + (OR)\pi_2} \quad (3.41)$$

(see Problem 2.4.1). Therefore, the power function for the test of two proportions, or the 2×2 table for two independent groups, is readily expressed in terms of relative risks or odds ratios.

It is then of interest to describe the factors that affect the power of a test to detect a given risk difference, relative risk, or odds ratio. This is easily done by considering the noncentrality factor $|\tau|$ in (3.39) as a function of the probabilities π_1 and π_2 in the two groups, as presented in Table 3.1 for the case of equal sample sizes. Examination of this table shows that for a fixed value of the risk difference $\Delta = \pi_1 - \pi_2$, as $\pi = (\pi_1 + \pi_2)/2$ approaches 0.5, then τ decreases, power $1 - \beta$ decreases, and thus the N required to achieve a given level of power increases. For example, consider the entries one step off the main diagonal that correspond to a fixed difference of $\Delta = 0.1$, with π_2 ranging from 0.1 to 0.8. The noncentrality parameter is a minimum when either probability equals 0.5. Therefore, the power to detect a fixed difference is greater when one of the outcomes is rare than when the two outcomes occur with probability close to 0.5. The reason is that the fixed difference

becomes proportionately smaller, and the variance larger, as the probability of either outcome approaches 0.5.

Even though the expressions for the noncentrality factor, sample size and power are stated in terms of the risk difference, it is common to describe the change in power as a function of the relative risk, or odds ratio. For a fixed value of the relative risk π_1/π_2 , the table shows that as π_2 increases, the total number of positive outcomes also increases, τ increases, power $1 - \beta$ increases, and the N required to provide a fixed level of power decreases. For example, compare the value of τ for $\pi_1/\pi_2 = 2$ and values of π_2 ranging from 0.1 to 0.4. Thus, the power to detect a given relative risk appears to depend directly on the expected number of outcomes. However, the reason is that the magnitude of the difference is growing as π_2 increases, and thus π_1 approaches 1 faster than π_2 approaches 0.5.

When the noncentrality parameters are examined in terms of odds ratios, a pattern similar to that for the risk difference is observed. For example, $OR = 6$ for values of $(\pi_1, \pi_2) = (0.4, 0.1), (0.6, 0.2), (0.8, 0.4)$, and $(0.9, 0.6)$. The respective values of the noncentrality parameter are 0.346, 0.408, 0.408, and 0.346.

Thus, when using power functions such as those in Figure 3.2 to evaluate the power over a range of parameters, the power curves using the risk difference, the relative risk or the odds ratio all depend on the value of π_2 . In each case, the actual values of the probabilities π_1 and π_2 determine power.

3.3.3 Equivalence

The preceding developments apply to the design of a study to establish superiority, or that the probability of the outcome with one treatment (or group) is greater than that of another. As described in Section 2.6.8, when a standard or positive control is employed as the basis for comparison, an alternative approach is to design a study to establish that the new therapy is equivalent, or noninferior, within a margin of error, to the standard. The evidence for equivalence can be obtained from a two-sided confidence interval or from the conduct of two one-sided tests. Makuch and Simon (1978) describe the evaluation of sample size from the perspective of a confidence interval; Blackwelder (1982) from one-sided tests.

Using the confidence interval approach, equivalence is established if the lower and upper $1 - \alpha$ confidence limits on the difference between groups fall within specified two-sided margins of error $(-\Delta_E, \Delta_E)$. Thus, the design specifies that the two-sided confidence limits on the difference between two proportions satisfy

$$(p_1 - p_2) \pm Z_{1-\alpha/2} \sigma_1 \in (-\Delta_E, \Delta_E), \quad (3.42)$$

where $\sigma_1^2 = \phi_1^2/N$, as in (3.32). We then wish to evaluate the probability that a study will satisfy this margin when indeed the new therapy is equivalent to the standard, or $\pi_1 = \pi_2 = \pi$, such that

$$p_1 - p_2 \stackrel{d}{\approx} \mathcal{N} [0, \sigma_0^2], \quad (3.43)$$

where $\sigma_0^2 = \phi_0^2/N$ as in (3.31). Thus, (3.42) is evaluated using σ_0 rather than σ_1 . Under these assumptions, with a given N , the probability that this margin would be

satisfied is

$$\begin{aligned} 1 - \beta &= P[(p_1 - p_2) \pm Z_{1-\alpha/2}\sigma_0 \in (-\Delta_E, \Delta_E) | H_0] \\ &= P[-\Delta_E + Z_{1-\alpha/2}\sigma_0 \leq (p_1 - p_2) \leq \Delta_E - Z_{1-\alpha/2}\sigma_0 | H_0] \\ &= \Phi(Z_{1-\beta/2}) - \Phi(-Z_{1-\beta/2}), \end{aligned} \quad (3.44)$$

where

$$Z_{1-\beta/2} = \frac{\Delta_E}{\sigma_0} - Z_{1-\alpha/2}. \quad (3.45)$$

The quantity $1 - \beta$ in this context is still called "power" since it is the probability of rejecting a tested hypothesis that a difference exists when using two simultaneous one-sided tests.

Then it follows that

$$\begin{aligned} \Delta_E &= (Z_{1-\alpha/2} + Z_{1-\beta/2})\sigma_0 \\ N &= \left[\frac{(Z_{1-\alpha/2} + Z_{1-\beta/2})\phi_0}{\Delta_E} \right]^2. \end{aligned} \quad (3.46)$$

To determine sample size or power for a specified margin Δ_E , one specifies the null probability π as in (3.35), from which ϕ_0 is computed. Both Makuch and Simon (1978) and Blackwelder (1982) present derivations in terms of a one-sided assessment only, as would be appropriate for an assessment of noninferiority. The Makuch-Simon equation is equivalent to the above. Blackwelder derives a similar expression but using the variance component ϕ_1 in lieu of ϕ_0 , where ϕ_1 is computed using specified probabilities (π_1, π_2) satisfying $\Delta_E = \pi_1 - \pi_2$.

Rather than evaluating (3.44) under the strict null hypothesis in (3.43), one might consider doing so under the hypothesis that the group probabilities differ to a small degree with the specified values π_1 and π_2 , $\pi_1 \neq \pi_2$. In this case, the distribution of $p_1 - p_2$ is shifted away from zero and the probability that the corresponding margin is satisfied decreases, leading to the need for a much larger N . For example, if $\pi_1 > \pi_2$, then the probability that the upper confidence limit is greater than Δ_E is increased, meaning that the probability that equivalence is demonstrated will decrease. If, indeed, there is the possibility that $\pi_1 > \pi_2$, then a noninferiority design would be more appropriate.

If the design is specified in terms of a relative risk or odds ratio, or the log of each, then the computation of sample size and power should be based on the distribution of the metric chosen. For a given value of π_2 , as shown in Problem 2.4, the value of π_1 yielding a specified value of the RR or OR is readily computed. However, if symmetric margins on the log RR or log OR are specified, then the corresponding margins on the risk difference are not symmetric. For example, if θ is the log odds ratio and a symmetric equivalence bound is specified with margins $(-\theta_E, \theta_E)$, then for a given π_2 , the difference $\pi_1 - \pi_2$ corresponding to $-\theta_E$ does not equal the negative of the difference $\pi_1 - \pi_2$ corresponding to θ_E (see Problem 3.6).

3.3.4 Noninferiority

The assessment of noninferiority requires that the specified upper or lower confidence limit on the difference between groups, $\pi_1 - \pi_2$, whichever is appropriate, at level $1 - \alpha$ falls within a margin of error. Assume that the groups are ordered such that the noninferiority specification is stated in terms of the upper confidence limit not exceeding a positive margin Δ_U . This would apply when group 1 is the new therapy, 2 the standard, and the outcome represents worsening, such that the probability of worsening with the new therapy does not substantially exceed that with the standard therapy. Alternatively, this would apply if group 1 is the standard, 2 the new, and the outcome represents improvement, so that the probability of improvement with the standard does not substantially exceed that of the new therapy.

The design then specifies that the one-sided confidence limit on the difference between two proportions not exceed a margin Δ_U for $\Delta_U > 0$, or that

$$(p_1 - p_2) + Z_{1-\alpha}\sigma_1 \leq \Delta_U. \quad (3.47)$$

If a two-sided confidence limit is specified, then $Z_{1-\alpha/2}$ is employed in this and subsequent expressions.

Under H_0 as in (3.43), with a given N , the probability that this margin would be satisfied is

$$P [(p_1 - p_2) + Z_{1-\alpha}\sigma_0 \leq \Delta_U \mid H_0] = 1 - \beta. \quad (3.48)$$

It follows that

$$Z_{1-\beta} = \frac{\Delta_U}{\sigma_0} - Z_{1-\alpha} \quad (3.49)$$

and

$$\begin{aligned} \Delta_U &= (Z_{1-\alpha} + Z_{1-\beta})\sigma_0 & (3.50) \\ N &= \left[\frac{(Z_{1-\alpha} + Z_{1-\beta})\phi_0}{\Delta_U} \right]^2. \end{aligned}$$

If the probability in (3.48) is evaluated under some other hypothesis with probabilities $\pi_{1(0)}, \pi_{2(0)}$, then $\Delta_U - (\pi_{1(0)} - \pi_{2(0)})$ is substituted for Δ_U , $\sigma_{(0)}^2$ for σ_0^2 , and $\phi_{(0)}^2$ for ϕ_0^2 in (3.49)–(3.50). The variance $\sigma_{(0)}^2$ is actually computed using each probability as in (3.32).

If the noninferiority design is specified in terms of an odds ratio or relative risk, then the required sample size could be computed by determining the probabilities π_1 and π_2 that correspond to a specified margin on the odds ratio or relative risk. In this case, a calculation based on the distribution of the log odds ratio or the log relative risk would be equivalent with large N (see Problem 3.6).

Example 3.3 Planning a Study for Equivalence or Noninferiority

Suppose that an established therapy provides treatment success (improvement) in 70% of patients ($\pi = 0.7$) and that it is desired to show that a new therapy is equivalent with a margin of $\Delta_E = 0.07$, two-sided. Assuming that the two therapies

are absolutely equivalent (i.e., $\pi_1 = \pi_2 = \pi = 0.7$), then from (3.46), with $\alpha = 0.05$, $N = 1982$ is required to provide 85% power that the equivalence margin is specified. If the response probabilities differed, then an even larger sample size would be required. For example, if it is assumed that the new therapy might indeed be slightly less effective, say $\pi_2 = 0.68$, compared to the standard with $\pi_1 = 0.70$, rather than H_0 , then direct evaluation of (3.44) with these probabilities over a range of sample sizes shows that $N = 3080$ would be needed to provide 85% power that equivalence would be established with the margin of 0.07.

An alternative design would be to establish noninferiority with a margin of $\Delta_U = 0.07$, that would represent the margin of superiority of the standard over the new therapy. Under H_0 using a one-sided 0.05-level confidence interval, $N = 1234$ provides 85% power that noninferiority would be established. If, however, it is assumed that the new therapy is actually slightly less effective with $\pi_{1(0)} = 0.7$ and $\pi_{2(0)} = 0.68$, the distribution of $p_1 - p_2$ is shifted toward the margin, requiring a larger sample size of $N = 2460$ to satisfy (3.47).

3.4 POWER OF CHI-SQUARE TESTS

3.4.1 Noncentral Chi-Square Distribution

The power of a chi-square test X_p^2 on p df is a monotonically increasing function of the noncentrality parameter, ψ^2 , of the noncentral distribution of the test statistic. This parameter is the expected value of the test under the alternative hypothesis, or $\psi^2 = E(X_p^2 | H_1)$, that can be factored as $\psi^2 = N\tau^2$, where τ^2 is the noncentrality factor. For a χ^2 test on p df at level α , the noncentrality parameter value that provides power $1 - \beta$ is denoted as $\psi^2(p, \alpha, \beta)$. For example, for a 1 df χ^2 test statistic, from (3.25), $\psi^2(1, \alpha, \beta) = (Z_{1-\alpha/2} + Z_{1-\beta})^2$. Thus, the noncentrality parameter that provides type II error probability $\beta = 0.1$ and power $1 - \beta = 0.9$ for a 1 df χ^2 test at $\alpha = 0.05$ is $\psi^2(1, 0.05, 0.10) = (1.96 + 1.645)^2 = 10.507$.

The SAS function CNONCT provides the value of the noncentrality parameter $\psi^2(p, \alpha, \beta)$ that yields the specified level of β for a test at level α on p df. The SAS function PROBCHI then provides the cumulative probabilities, and CINV the quantiles, of the chi-square distribution under a noncentral distribution.

To determine sample size using this approach, one first obtains the value $\psi^2(p, \alpha, \beta)$ of the noncentrality parameter that will provide the desired level of power for the noncentral chi-square distribution. The value of the noncentrality factor τ^2 under the alternative hypothesis is then specified or evaluated under an appropriate model, and is usually defined using the variance under the null hypothesis. Given the value of τ^2 , the N required to provide power $1 - \beta$ is that value for which $\psi^2(p, \alpha, \beta) = N\tau^2$, yielding

$$N = \frac{\psi^2(p, \alpha, \beta)}{\tau^2}. \quad (3.51)$$

For a 1 df χ^2 or a two-sided Z -test, this is equivalent to the simplifications presented in (3.22)–(3.27).

In some cases, these expressions are obtained under a local alternative in which case the calculations describe the limiting power function of the test. These concepts are described in Section 3.5.

3.4.2 Pearson Chi-Square Tests

Meng and Chapman (1966) described the limiting power function of chi-square tests for $R \times C$ contingency tables. These results were employed by Lachin (1977) and Guenther (1977) to describe the power function and to determine the sample size needed to provide a desired level of power for the contingency chi-square test for an $R \times C$ table.

3.4.2.1 A Multinomial Distribution Consider the case of C mutually exclusive categories (cells) with frequencies $\{n_j\}$ distributed as multinomial with probabilities $\{\pi_j\}$ ($j = 1, \dots, C$) that can be expressed as functions $\pi_j(\boldsymbol{\theta})$ of q underlying parameters $\boldsymbol{\theta} = (\theta_1 \dots \theta_q)^T$. A null hypothesis stated in terms of the $\boldsymbol{\theta}$, $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$, yields corresponding multinomial probabilities $\pi_{0j} = \pi_j(\boldsymbol{\theta}_0)$. For the case where the complete vector $\boldsymbol{\theta}_0$ is specified, the expected frequencies are $E_j = N\pi_{0j}$ and the Pearson goodness-of-fit test X_{P,π_0}^2 presented in (2.122) is distributed as chi-square on $C-1$ *df*. Under an alternative hypothesis, $H_1: \boldsymbol{\theta} = \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$, from Slutsky's theorem (A.45), the limiting expectation of X_{P,π_0}^2 is the noncentrality parameter $\psi^2 = N\tau^2$, where

$$\tau^2 = \sum_{j=1}^C \frac{[\pi_j(\boldsymbol{\theta}_1) - \pi_j(\boldsymbol{\theta}_0)]^2}{\pi_j(\boldsymbol{\theta}_0)}. \quad (3.52)$$

When the q parameters $\boldsymbol{\theta}_0$ are estimated from the sample, $\hat{E}_j = N\pi_j(\hat{\boldsymbol{\theta}}_0)$ and the Pearson test X_P^2 in (2.125) is distributed as noncentral chi-square on $C - q - 1$ *df* with noncentrality parameter as in (3.52).

Example 3.4 Recombination Fraction

Consider a test of the hypothesis of independent Mendelian segregation of two traits, as in Example 2.15, where under H_0 the recombination fraction is $\theta = \theta_0 = 0.25$. Suppose that we wish to detect $H_1: \theta = \theta_1 = 0.4$ under the alternative. From (2.126), the probabilities of the four phenotypes $\{\pi_j(\boldsymbol{\theta})\}$ under H_0 and H_1 are

	<i>ab</i>	<i>Ab</i>	<i>aB</i>	<i>AB</i>
H_0	1/16	3/16	3/16	9/16
H_1	0.1	0.15	0.15	0.6

In the test of goodness of fit, all cell probabilities are obtained from the specified value for $\boldsymbol{\theta}$ (i.e., none are estimated from the data), and thus the test is distributed as chi-square on 3 *df*. The test requires a noncentrality parameter value of $\psi^2(3, 0.05, 10) = 14.1715$ to provide 90% power. The resulting value of the noncentrality factor in (3.52) is $\tau^2 = 0.04$, that results in a required $N = 355$.

3.4.2.2 $R \times C$ Contingency Table For an $R \times C$ contingency table with cell frequencies $\{n_{ij}\}$ and corresponding probabilities $\{\pi_{ij}\}$, the estimated expected frequencies $\widehat{E}_{ij} = n_{i\bullet}n_{\bullet j}/N$ can either be derived under the null hypotheses of homogeneity among the R groups, or equivalently, of statistical independence between the row and column classifications. The resulting contingency test X_P^2 presented in (2.143) is distributed as chi-square on $(R - 1)(C - 1)$ df.

The probabilities under the null hypothesis of independence are specified as $\pi_{0ij} = \pi_{i\bullet}\pi_{\bullet j}$, and those under the alternative as π_{ij} , where $\pi_{ij} \neq \pi_{0ij}$ for some cells (ij) , where the two sets of marginal probabilities are equal; i.e., $\sum_{i=1}^R \pi_{ij} - \pi_{0ij} = 0$ for every j and $\sum_{j=1}^C \pi_{ij} - \pi_{0ij} = 0$ for every i . Then the limiting expected value of the chi-square test statistic under the alternative hypothesis is $E(X_P^2) = N\tau^2$, where

$$\tau^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{[\pi_{ij} - \pi_{0ij}]^2}{\pi_{0ij}}. \quad (3.53)$$

The probabilities under H_0 and H_1 could be obtained by specifying the marginal probabilities $\{\pi_{i\bullet}\}$ and $\{\pi_{\bullet j}\}$ that provide the table of values $\{\pi_{0ij}\}$. Then the cell probabilities $\{\pi_{ij}\}$ under the alternative hypothesis are specified, yielding deviations from the null for each cell, $\delta_{ij} = \pi_{ij} - \pi_{0ij}$, that sum to zero within each row and column. The resulting noncentrality factor is

$$\tau^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{[\delta_{ij}]^2}{\pi_{i\bullet}\pi_{\bullet j}}. \quad (3.54)$$

Often, however, it is easier to specify these probabilities under the hypothesis of homogeneity among the R groups. As in Section 2.11.1, let η_i denote the probability of falling in the i th group, where in the above notation $\eta_i = \pi_{i\bullet}$. Then let $\boldsymbol{\pi}_i = (\pi_{1(i)} \cdots \pi_{C(i)})^T$ denote the vector of conditional probabilities of the C response categories within the i th group, where in the above notation $\pi_{ij} = \pi_{j(i)}\eta_i$. Under the hypothesis of homogeneity, $\boldsymbol{\pi}_1 = \cdots = \boldsymbol{\pi}_R = \boldsymbol{\pi} = (\pi_{\bullet 1} \cdots \pi_{\bullet C})^T$. As in the test of two proportions in Section 3.3, the vector of probabilities under the null is obtained as a weighted average of those under the alternative, weighted proportionally to the sample fractions. Thus, once the vectors $\{\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_R\}$ are specified, the elements of $\boldsymbol{\pi}$ are obtained as $\boldsymbol{\pi} = \sum_{i=1}^R \eta_i \boldsymbol{\pi}_i$. This yields a matrix of deviations from homogeneity $\delta_{ij} = \pi_{j(i)} - \pi_{\bullet j}$, where $\delta_{ij} \neq 0$ for some cells (ij) . Then, Lachin (1977) shows that the noncentrality factor is

$$\tau^2 = \sum_{j=1}^C \frac{\sum_{i=1}^R \eta_i \delta_{ij}^2}{\pi_{\bullet j}}. \quad (3.55)$$

For a 2×2 table, it is readily shown that $E(X^2) = N\tau^2$ for τ , as shown in (3.39).

Note that if the design will employ equal sample fractions, $\eta_i = 1/R$, then the vector $\boldsymbol{\pi}$ could be specified directly and the matrix of deviations $\{\delta_{ij}\}$ within each group so that $\pi_{j(i)} = \pi_{\bullet j} + \delta_{ij}$.

Example 3.5 A 3×3 Table

For example, let the following be the pattern of conditional probabilities $\{\pi_{\bullet j}\}$ expected under H_0 for a planned clinical trial comparing three equal-sized treatment groups ($\eta_i = 1/3$) with respect to three categories of response (A, B, C , corresponding to $j = 1, 2, 3$) with probabilities under the null hypothesis:

$$\{\pi_{\bullet j}\} = \begin{array}{ccc} A & B & C \\ 0.10 & 0.15 & 0.75 \end{array}$$

Under the alternative, assume that we wish to detect the following pattern of differences $\{\delta_{ij}\}$ among the three groups, where $\pi_{j(i)} = \pi_{\bullet j} + \delta_{ij}$:

Group (i)	A	B	C
1	-0.09	-0.11	0.20
2	-0.01	0.01	0
3	0.10	0.10	-0.20

and $\sum_{i=1}^R \delta_{ij} = 0$ for all $j = 1, 2, 3$. Substituting these values into (3.55) yields $\tau^2 = 0.1456$. The SAS function CINV(0.95, 4) provides the critical value at $\alpha = 0.05$ with 4 df of 9.488. The SAS function CNONCT(9.488, 4, 0.10) provides the value of the noncentrality parameter $\psi^2(4, 0.05, 0.10) = 15.405$, that yields $\beta = 0.10$ for a test on 4 df at $\alpha = 0.05$. Solving for $N = \psi^2/\tau^2$ yields $N = 106$.

To determine power based on a given sample size N , one simply determines the value of the noncentrality parameter $\psi^2 = N\tau^2$ and then evaluates the cumulative probability at the critical value under the noncentral chi-square distribution. For the above example, if $N = 80$, then $\psi^2 = 80(0.1456) = 11.48$. The SAS function PROBCHI(9.488, 4, 11.48) then provides $\beta = 0.2112$ and power $1 - \beta = 0.7888$.

If there are unequal group sample sizes, the values under the null are a weighted average of the $\{\pi_{j(i)}\}$ under the alternative, in which case it is most convenient to start with the latter. Then the values under the null are computed as $\pi_{\bullet j} = \sum_{i=1}^R \eta_i \pi_{j(i)}$, and the $\{\delta_{ij}\}$ are obtained by subtraction.

3.4.3 The Mean Score (Rank) Test

The Cochran-Mantel-Haenszel mean score test with rank scores is equivalent to a Wilcoxon rank sum test for two groups, or the Kruskal-Wallis test for $R > 2$ groups. The Wilcoxon test is also algebraically equal to the Mann-Whitney test (Mann and Whitney, 1947) of the Mann-Whitney proversion or competing probability, say ϕ . If Y_1 and Y_2 denote a strictly continuous response variable in each group (i.e., with no ties), then the Mann-Whitney parameter is $\phi = P(Y_1 > Y_2)$, where $\phi = 0.5$ under H_0 : $f(y_1) = f(y_2)$ with common density $f(\cdot)$. Lehmann (1975) and Noether (1987) describe the power function of the test against such alternatives, assuming no ties.

For a discrete response variable, the test can be framed in terms of the parameter $\phi = P(Y_1 > Y_2) + 0.5P(Y_1 = Y_2)$, again with expectation 0.5 under H_0 . For a variable with C categories, O'Brien and Castelloe (2006) describe the power of

a test of the Mann-Whitney log odds $\log[\phi/(1 - \phi)]$ for two specified vectors of probabilities among groups, allowing for ties, and this method is implemented in the SAS procedure POWER. Zhao et al. (2008) present an alternative expression for the power of a test of ϕ itself, allowing for ties using an approximate expression for the variance of the estimate under the null. The power of the Wilcoxon-Mann-Whitney test has also been described under specific models, such as the proportional odds model.

However, simple expressions for the power of the mean score test are readily derived from the noncentrality parameter of the mean score test itself. In this approach the rank scores are assumed fixed and the frequencies within each group are assumed to be randomly distributed under the respective multinomial distributions. Thus, the expected rank score is provided by the expected proportions in each category.

First consider the case where numerical table scores s_1, \dots, s_C are associated with each category. Then, from Slutsky's theorem, the limiting expectation of (2.140) is $E[X_s^2] = N\tau^2$, where

$$\tau^2 = \frac{[s'(\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)]^2}{s'\boldsymbol{\Sigma}(\boldsymbol{\pi})s \left[\frac{1}{\eta_1} + \frac{1}{\eta_2} \right]} \quad (3.56)$$

with sample fractions η_1 and η_2 in the two groups. For specific vectors $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ in the two groups, $\boldsymbol{\pi} = \eta_1\boldsymbol{\pi}_1 + \eta_2\boldsymbol{\pi}_2$. Sample size or power can then be computed from the noncentral chi-square distribution.

For rank scores, the vector of expected scores $E(\mathbf{s})$ is a function of the average vector of response probabilities $\boldsymbol{\pi} = (\pi_{\bullet 1} \cdots \pi_{\bullet C})^T$, that is assumed to be the same under either H_0 or H_1 . Then, from (2.137),

$$\begin{aligned} E(s_1) &= N\pi_{\bullet 1}/2 \\ E(s_j) &= N \left[\sum_{\ell=1}^{j-1} \pi_{\bullet \ell} + \pi_{\bullet j}/2 \right], \quad j > 1. \end{aligned} \quad (3.57)$$

Denoting $E(s_j) = Nv_j$ and $E(\mathbf{s}) = N\mathbf{v}$, substituting into (3.56) yields

$$\tau^2 = \frac{[\mathbf{v}'(\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)]^2}{\mathbf{v}'\boldsymbol{\Sigma}(\boldsymbol{\pi})\mathbf{v} \left[\frac{1}{\eta_1} + \frac{1}{\eta_2} \right]} \quad (3.58)$$

where τ^2 is uniquely defined from the vectors $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ and the sample fractions η_1 and η_2 .

Example 3.6 *Albuminuria in the DCCT*

As described in Section 1.5, an outcome of the DCCT was the development of microalbuminuria, defined as an albumin excretion rate ≥ 40 mg/24 h. This represents an abnormal value relative to the normal population, but one below the clinical level of macroalbuminuria, defined as a value ≥ 300 mg/24 h, at which nephropathy is clinically detectable. Thus, it would be of interest to assess the sample size requirement and power of a mean score test for the difference between the intensive and

conventional groups for this outcome at five years of follow-up. The following table represents an alternative hypothesis of interest with conditional probabilities π_1 and π_2 within the two groups, respectively

Group	Albuminuria		
	Normal	Micro	Macro
Conventional (π_1)	0.85	0.10	0.05
Intensive (π_2)	0.90	0.075	0.025

With equal sample sizes, the values under the null are assumed to be $\pi = (0.875 \ 0.0875 \ 0.0375)^T$. The noncentrality parameter required to provide 90% power to detect the above difference using a two-sized test at the 0.05 level on 1 *df* is $\psi^2(1, 0.05, 0.10) = 10.507$. For this alternative hypothesis, from (3.56), a mean score test with the table scores $s = (1 \ 2 \ 3)^T$ yields $\tau^2 = 0.00666$, such that a sample size $N = 1578$ is required. Using rank scores, then from (3.58), $\tau^2 = 0.00598$ and $N = 1757$. The larger sample size with the rank scores indicates that the table scores test would have greater power than the rank scores test.

However, in the DCCT, only about $N = 600$ were expected to be evaluated after five years of follow-up in the secondary cohort. This provides a noncentrality parameter of $\psi^2 = (600)(0.00666) = 3.59$. The realized power is then obtained using the statement "power = 1 - probchi(3.84, 1, 3.59);", where $\chi^2_{1-\alpha} = 3.84$ is the critical value on 1 *df* at $\alpha = 0.05$. The test would provide power = 0.474.

Numerically, the above calculations are identical to those provided by the equation of Zhao et al. (2008). They are also close to those provided by the SAS procedure POWER (see Section 3.5).

3.4.4 The Cochran-Armitage Test of Trend

The Cochran-Armitage test for trend X_{CA}^2 in (2.149) assumes that the probability of a binary response π_i in the *i*th of *R* groups is a linear function of the monotonically increasing or decreasing score s_i associated with the *i*th group of sample size $N\eta_i$ with sample fraction η_i . With table scores, the s_i are simply the numerical values associated with each group, such as in a dose ranging study where s_i is the dose of the drug administered to the *i*th group. Then the noncentrality factor is

$$\tau^2 = \frac{\left[\sum_{i=1}^R \eta_i s_i (\pi_i - \pi) \right]^2}{\pi(1-\pi) \sum_{i=1}^R \eta_i (s_i - \bar{s})^2} = \frac{\left[\sum_{i=1}^R \eta_i \pi_i (s_i - \bar{s}) \right]^2}{\pi(1-\pi) \sum_{i=1}^R \eta_i (s_i - \bar{s})^2}, \quad (3.59)$$

where $\bar{s} = \sum_{i=1}^R \eta_i s_i$ and $\pi = \sum_{i=1}^R \eta_i \pi_i$. Sample size and power are then obtained directly using the noncentral chi-square distribution. Slager and Schaid (2001) present alternative expressions as a function of the *R* multinomial parameters for the positive frequencies among the groups, and those for the *R* negative frequencies. With the null hypothesis variance, their expression for a two-sided test equals the noncentrality factor above.

This expression requires specification of the $\{\pi_i\}$, $\{\eta_i\}$, and $\{s_i\}$. The test itself, however, is derived as simply $\widehat{\beta}^2/\widehat{V}(\widehat{\beta})$, where the model specifies that

$$\pi_i = \alpha + \beta s_i = \pi + \beta(s_i - \bar{s}) \quad (3.60)$$

and thus

$$\sum_{i=1}^R \eta_i s_i (\pi_i - \pi) = \beta \sum_{i=1}^R \eta_i s_i (s_i - \bar{s}) = \beta \sum_{i=1}^R \eta_i (s_i - \bar{s})^2. \quad (3.61)$$

Then

$$\tau^2 = \frac{\beta^2 \sum_{i=1}^R \eta_i (s_i - \bar{s})^2}{\pi(1 - \pi)}, \quad (3.62)$$

that is readily evaluated for any β with a specified set of scores $\{s_i\}$ and group sample fractions $\{\eta_i\}$.

Alternatively, as in Example 2.20, rank scores may be employed, in which case the expected fractional rank scores $\{v_i\}$ are obtained in a manner similar to that in (3.57) but as a function of the sample fractions η_i ; i.e., where $v_1 = \eta_1/2$ and $v_i = \sum_{\ell=1}^{i-1} \eta_\ell + \eta_i/2$ for $i > 1$. The noncentrality factor using the test in (2.152) is

$$\tau^2 = \frac{\beta_f^2 \sum_{i=1}^R \eta_i (v_i - \bar{v})^2}{\pi(1 - \pi)}, \quad (3.63)$$

where β_f is the change in π over the range of values because a unit change in v represents the entire range. With equal sample fractions ($\eta_i = 1/R$), the noncentrality parameter for the test in (2.156) further simplifies to

$$\psi^2 = \frac{\beta_f^2 n (R + 1) (R - 1)}{12R\pi(1 - \pi)}. \quad (3.64)$$

When formulated in terms of the coefficient per unit change in group (β_g) in (2.156), then

$$\psi^2 = \frac{\beta_g^2 n (R + 1) R (R - 1)}{12\pi(1 - \pi)}. \quad (3.65)$$

In both cases $\psi^2 = n\tau^2$, so that the solution for sample size is computed in terms of the sample size per group.

Example 3.7 Genetic Marker

The test of trend is commonly used to assess the association between the prevalence of a trait and the number of high- or low-risk alleles of a candidate gene in a case-control study. Label the high-risk allele as A , the low-risk allele as a . Under the null hypothesis of Hardy-Weinberg equilibrium, the proportion of cases should have no association with the number of high-risk alleles whereas under disequilibrium there should be an increasing proportion of cases as the number of high-risk alleles increases, from 0 (aa) to 1 (Aa) to 2 (AA). Slager and Schaid (2001) present an example of a case-control study with equal numbers of cases and controls (the binary

responses); i.e., $\pi = 0.5$, where under the alternative the probabilities of being a case within each category are 0.473, 0.647, and 0.818 under a specific alternative. Using the above expression for table scores in (3.59) yields $\tau^2 = 0.0162$. For a test on 1 df at $\alpha = 0.05$ with critical value 3.842, the noncentrality parameter required to provide 90% power is $\psi^2 = 10.51$. The resulting sample size required is $N = 651$.

Example 3.8 Coronary Heart Disease in the Framingham Study

Example 2.20 presents a test of trend in the prevalence of coronary heart disease as a function of the level of serum cholesterol, grouped into four categories, in the Framingham study. Suppose that we wished to conduct a confirmatory study using cholesterol grouped by quartiles (i.e., $\eta_i = 1/4$) and where the probabilities of the outcome over quartiles are $\{\pi_i\} = (0.03, 0.04, 0.08, 0.14)$. The rank scores then are $\{s_i\} = (0.125, 0.375, 0.625, 0.875)$. The resulting $\beta_f = 0.148$. Substituting into (3.64) yields $\tau^2 = 0.10179$. Then for 90% power as above, $\psi^2 = 10.51$ and the required sample size per group is 103.22, that yields a total sample size $N = 413$. The corresponding slope per group number is $\beta_g = 0.037$.

3.5 SAS PROC POWER

The SAS procedure POWER provides calculation of sample size and power for one- and two-sample analyses of frequencies, means, ranks, and survival outcomes. For illustration, computations for the test for two proportions and the Wilcoxon test are described.

3.5.1 Test for Two Proportions

Example 3.2 describes computations for planning a study to detect a relative risk of 0.70 when the control probability is 0.40. The computation of sample size could be obtained using the following statements

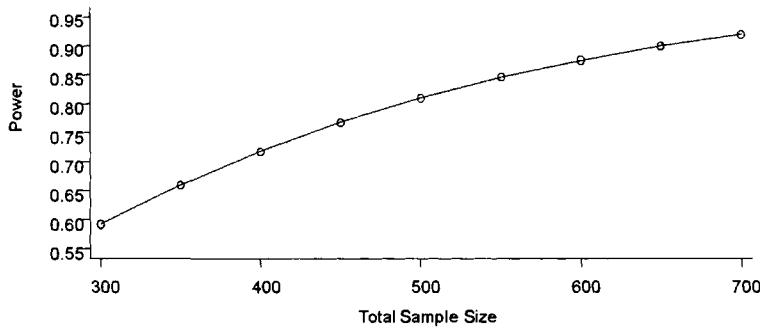
```
proc power;
twosamplefreq test=pchi alpha = 0.05 power = 0.90 NTOTAL = .
refproportion = 0.40 relativisk = 0.70;
```

This yields $N = 652$, as shown in the Example 3.2.

The procedure can also compute power for a given N , such as $N = 500$, by changing the options to read "NTOTAL = 500 power = ." The following statements would do so for other values of N and generate a plot of the power function of the test:

```
proc power;
twosamplefreq test=pchi alpha = 0.05 power = . NTOTAL = 300 500
700 refproportion = 0.40 relativisk = 0.70; plot step=50;
```

Fig. 3.3 Power over a range of sample sizes for the test of two proportions to detect a relative risk of 0.7 with a control group probability of 0.4 with two equal-sized groups with $\alpha = 0.05$, two-sided.



The resulting power function is displayed in Figure 3.3. For a sample size of 500, the estimated power is 0.81 as in Example 3.2.

3.5.2 Wilcoxon Mann-Whitney Test

For the two-sample Wilcoxon test, POWER employs the method of O'Brien and Castelloe (2006), that is based on the distribution of the log odds of the Mann-Whitney parameter computed from the distribution of rank scores under the specified alternative. The method described in Section 3.4.3, however, is based on the specified vectors of probabilities from which the expected rank scores are obtained. The two methods produce equivalent results. To employ POWER for the above example, the following statements would be used:

```
proc power;
twosamplewilcoxon alpha = 0.05 power = 0.9 NTOTAL = .
vardist("group1") = ordinal ((1 2 3) : (0.85 0.10 0.05))
vardist("group2") = ordinal ((1 2 3) : (0.90 0.075 0.025))
variables = "group1" | "group2";
```

The program provides $N = 1754$ that is negligibly different from the value 1757 provided by the noncentral distribution of the test statistic.

Alternatively, the following statements would provide the computation of power of the test for other sample sizes:

```
proc power; twosamplewilcoxon alpha = 0.05 NTOTAL = 500 600 1700
power = .
vardist.....(etc.);
```

For a sample size of 600, the estimated power is 0.475, equivalent to that computed above from the noncentral distribution.

3.6 EFFICIENCY

3.6.1 Pitman Efficiency

Generalizations of the expression for the power function of a test may also be used to describe the asymptotic efficiency of a test statistic under a family of local alternatives, an approach originally due to Pitman (1948) and elaborated by Noether (1955). This approach is important when a variety of competing statistics could be used to test a particular null versus alternative hypothesis, in which case it is desirable to employ whichever test is most powerful. However, many such families of tests are based on asymptotic distribution theory, and asymptotically any test in the family will provide a level of power approaching 1.0 against any alternative hypothesis away from the null hypothesis. Thus, Pitman introduced the concept of the asymptotic power (efficiency) against a family of *local alternatives* that remain "close" to the null hypothesis value for large values of N , so that the power remains below 1.0 except in the limiting case ($n \rightarrow \infty$).

Let T be a test statistic for $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$, where the power function of T increases in θ , so that T is an unbiased test, and where the power also increases in the total sample size n , so that T is also consistent (cf. Lehmann, 1986). Let $\alpha_n(\theta_0)$ designate the Type I error probability of the test for a given sample size n , and let $\gamma_n(\theta) = 1 - \beta_n(\theta)$ designate the power of the test to detect a specific value $\theta \neq \theta_0$ for a given n as n increases indefinitely. Asymptotically, for any such test, $\lim_{n \rightarrow \infty} \gamma_n(\theta) \rightarrow 1$ for any $\theta \neq \theta_0$ and thus the limiting power in a broad sense is not useful for evaluating any particular test, nor for the comparison of different tests at the same value of θ under H_1 . Therefore, we use the concept of *local power* defined under a sequence of local alternatives as n increases of the form

$$\theta_n = \theta_0 + \frac{\delta}{\sqrt{n}}, \quad (3.66)$$

such that $\lim_{n \rightarrow \infty} \theta_n = \theta_0$ for any finite δ . Since $Z_{1-\beta} = Z_{\gamma_n}$ increases in \sqrt{n} , then Z_{γ_n} approaches a constant at the same rate that $\theta_n \rightarrow \theta_0$ so that $\alpha < \gamma_n(\theta_n) < 1$.

For such a sequence of local alternatives, the efficiency of the test is a function of the limiting power

$$\lim_{n \rightarrow \infty} \gamma_n(\theta_n) = \lim_{n \rightarrow \infty} 1 - \beta_n(\theta_n), \quad (3.67)$$

or

$$\lim_{n \rightarrow \infty} P [T_n > T_{n,\alpha} \mid \theta_n], \quad (3.68)$$

where $T_{n,\alpha}$ is the α -level critical value. If we restrict consideration to test statistics that are asymptotically normally distributed, under the alternative hypothesis the asymptotic distribution can be described as

$$T_n \mid \theta_n \xrightarrow{d} N [\mu(\theta_n), \phi^2(\theta_n) / n] \quad (3.69)$$

with mean $E(T_n \mid \theta_n) = \mu(\theta_n)$ and variance $V(T_n \mid \theta_n) = \sigma^2(\theta_n)$ that may depend on the value of θ_n with variance component $\phi^2(\theta_n)$. Then under $H_0: \theta = \theta_0$, asymptotically

$$Z_n = \frac{T_n - \mu(\theta_0)}{\phi(\theta_0) / \sqrt{n}} \xrightarrow{d} N(0, 1). \quad (3.70)$$

Therefore, for an upper-tailed $(1 - \alpha)$ -level test,

$$\gamma_n(\theta_n) = P [Z_n \geq Z_{1-\alpha} \mid \theta_n]. \quad (3.71)$$

From the description of the power function of such a test (3.20), where $1 - \beta_n(\theta_n) = \gamma_n(\theta_n) = \Phi[Z_{\gamma_n(\theta_n)}]$, yields

$$Z_{\gamma_n(\theta_n)} = \frac{\mu(\theta_n) - \mu(\theta_0)}{\phi(\theta_n) / \sqrt{n}} - Z_{1-\alpha} \frac{\phi(\theta_0)}{\phi(\theta_n)}. \quad (3.72)$$

Now assume that $\mu(\theta_n)$ and $\phi(\theta_n)$ are continuous and differentiable at $\theta = \theta_0$. Evaluating γ_n under a sequence of local alternatives, the limiting power is provided by $\lim_{n \rightarrow \infty} Z_{\gamma_n(\theta_n)}$. For the second term on the right hand side, evaluating $\phi(\theta_n)$ under the local alternative yields

$$\lim_{n \rightarrow \infty} \frac{Z_{1-\alpha} \phi(\theta_0)}{\phi\left(\theta_0 + \frac{\delta}{\sqrt{n}}\right)} \rightarrow Z_{1-\alpha}. \quad (3.73)$$

Therefore, the limiting value of the noncentrality parameter for the test is

$$\lim_{n \rightarrow \infty} [Z_{\gamma_n(\theta_n)}] + Z_{1-\alpha} = \lim_{n \rightarrow \infty} \left[\frac{\mu(\theta_n) - \mu(\theta_0)}{\phi(\theta_0) / \sqrt{n}} \right]. \quad (3.74)$$

Multiplying the numerator and denominator by δ , then

$$\lim_{n \rightarrow \infty} [Z_{\gamma_n(\theta_n)}] + Z_{1-\alpha} = \frac{\delta}{\phi(\theta_0)} \lim_{n \rightarrow \infty} \left[\frac{\mu\left(\theta_0 + \frac{\delta}{\sqrt{n}}\right) - \mu(\theta_0)}{\delta / \sqrt{n}} \right]. \quad (3.75)$$

Now let $\varepsilon = \delta / \sqrt{n}$, so that $\lim_{n \rightarrow \infty} H(\delta / \sqrt{n}) \equiv \lim_{\varepsilon \rightarrow 0} H(\varepsilon)$ for any function $H(\cdot)$. Then

$$\begin{aligned} \lim_{n \rightarrow \infty} [Z_{\gamma_n(\theta_n)}] + Z_{1-\alpha} &= \frac{\delta}{\phi(\theta_0)} \lim_{\varepsilon \rightarrow 0} \left[\frac{\mu(\theta_0 + \varepsilon) - \mu(\theta_0)}{\varepsilon} \right] \\ &= \frac{\delta}{\phi(\theta_0)} \left(\frac{d\mu(\theta_0)}{d\theta_0} \right) = \frac{\delta \mu'(\theta_0)}{\phi(\theta_0)}. \end{aligned} \quad (3.76)$$

Since δ is a constant by construction, then the efficiency of any test at level α will be proportional to

$$\lim_{n \rightarrow \infty} [Z_{\gamma_n(\theta_n)}] \propto \frac{\mu'(\theta_0)}{\phi(\theta_0)} \propto \frac{[\mu'(\theta_0)]^2}{\phi^2(\theta_0)}. \quad (3.77)$$

Therefore, the efficiency of any given statistic T as a test of H_0 versus H_1 is usually defined as

$$Eff(T) = \left[\left(\frac{dE(T)}{d\theta} \right)^2 \bigg/ \phi^2(\theta) \right]_{\theta=\theta_0}. \quad (3.78)$$

The numerator is the derivative of the expectation of T evaluated at the null hypothesis value. When $\mu'(\theta) = 1$, such as when T is unbiased or consistent for θ itself such that $E(T) = \mu(\theta) = \theta$, then

$$Eff(T) = \frac{1}{\phi^2(\theta_0)} \propto \frac{1}{\phi^2(\theta_0)/n} = \frac{1}{V(T|\theta_0)}, \quad (3.79)$$

which is the reciprocal of the large sample variance of T under the null hypothesis (the null variance).

3.6.2 Asymptotic Relative Efficiency

If there are two alternative tests for $H_0: \theta = \theta_0$, say T_1 and T_2 , then the locally more powerful test of the two will be that for which the efficiency is greatest. The difference in efficiency is reflected by the *asymptotic relative efficiency* (*ARE*) of T_1 to T_2 :

$$ARE(T_1, T_2) = \frac{Eff(T_1)}{Eff(T_2)}. \quad (3.80)$$

A value of $ARE(T_1, T_2) > 1$ indicates that T_1 is the more powerful test asymptotically, whereas $ARE(T_1, T_2) < 1$ indicates that T_2 is more powerful.

If T_1 and T_2 are both unbiased such that $E(T_i) = \theta$ for $i = 1, 2$, then

$$ARE(T_1, T_2) = \frac{[\phi_{T_1}^2(\theta_0)]^{-1}}{[\phi_{T_2}^2(\theta_0)]^{-1}} = \frac{\sigma_{T_2}^2(\theta_0)}{\sigma_{T_1}^2(\theta_0)} = \frac{V(T_2|\theta_0)}{V(T_1|\theta_0)},$$

where T_1 is the more powerful test when $\sigma_{T_1}^2 < \sigma_{T_2}^2$, and vice versa. From (3.21), assuming that $\sigma_0^2 \doteq \sigma_1^2 \doteq \sigma^2$, the sample size required to provide power $1 - \beta$ to detect a difference Δ using a test at level α is of the form

$$N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \phi^2}{\Delta^2} \propto \phi^2, \quad (3.81)$$

which is directly proportional to the variance of the statistic. Thus, for such tests, the $ARE(T_1, T_2)$ can be interpreted in terms of the relative sample sizes required to provide a given level of power against a sequence of local alternatives:

$$ARE(T_1, T_2) \doteq \frac{N(T_2)}{N(T_1)}, \quad (3.82)$$

where $N(T_i)$ is the sample size required to achieve power $1 - \beta$ for a fixed Δ with test T_i .

Example 3.9 ARE of Normal Median: Mean

Consider sampling from a normal distribution, $X \sim \mathcal{N}(\mu, \phi^2)$, where we wish to test $H_0: \mu = 0$ versus $H_1: \mu \neq 0$. Two consistent estimators of μ are

$$T_1 = \bar{x} \sim \mathcal{N}(\mu, \phi^2/n) \quad (3.83)$$

$$T_2 = x_{(0.5n)} = \text{median} \sim \mathcal{N}[\mu, (\pi\phi^2)/(2n)], \quad (3.84)$$

where $\pi/2 = 1.57$. Thus,

$$ARE(T_1, T_2) = \frac{[\phi^2/n]^{-1}}{[(1.57)\phi^2/n]^{-1}} = 1.57. \quad (3.85)$$

Therefore, to provide the same level of power against a sequence of local alternatives, the test using the median would require a 57% greater sample size than the test using the mean. Thus, the test based on the mean is 57% more efficient than the test using the median.

3.6.3 Estimation Efficiency

An equivalent expression for the efficiency of a statistic may also be obtained from the perspective of the estimation efficiency of that statistic as an estimate of the parameter of interest. Let $\hat{\theta}$ be a consistent estimator of θ , where $\hat{\theta}$ is a function of a sample statistic T of the form $\hat{\theta} = f(T)$. Then, from the law of large numbers and Slutsky's theorem (A.47), as $n \rightarrow \infty$, then $T \xrightarrow{p} \tau = E(T)$ and $\hat{\theta} \xrightarrow{p} \theta = f(\tau)$. Therefore, by a Taylor's expansion of $f(T)$ about τ , asymptotically

$$\hat{\theta} \cong \theta + (T - \tau) \left[\frac{df(T)}{dT} \right]_{T=\tau} \quad (3.86)$$

and

$$(\hat{\theta} - \theta) \cong \frac{T - \tau}{\left(\frac{dE(T)}{d\theta} \right)}. \quad (3.87)$$

Therefore, asymptotically

$$V(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \cong \frac{E(T - \tau)^2}{\left(\frac{dE(T)}{d\theta} \right)^2} \cong \frac{V(T)}{\left(\frac{dE(T)}{d\theta} \right)^2} = \frac{1}{Eff(T)}. \quad (3.88)$$

Thus, $V(\hat{\theta})^{-1}$ is equivalent to the Pitman efficiency of the test statistic T in (3.77) evaluated at the true value of θ .

In some cases these concepts are used to evaluate the properties of a set of tests within a specific family that may not be all-inclusive. For example, if the estimator T is defined as a linear combination of other random variables, then we refer to the efficiency of the family of linear estimators. In this case, another test that is not a member of this family, such as a nonlinear combination of the random variables, may have greater efficiency. The minimum variance, and thus maximum efficiency, for any estimation-based test is then provided by the Cramér -Rao lower bound for the variance of the estimate as given in (A.109):

$$V(T | \theta) \geq \frac{[\mu'_T(\theta)]^2}{I(\theta)}, \quad (3.89)$$

where $E(T|\theta) = \mu_T(\theta)$ is some function of θ and $I(\theta)$ is the information function derived from the likelihood function for θ (see Section A.6.4). For an unbiased estimator of θ , the maximum possible efficiency of the estimator or test is $I(\theta)$.

These concepts of efficiency and asymptotic relative efficiency, in general, are widely used in the study of distribution-free statistics. Hájek and Šídák (1967) have developed general results that provide the asymptotically most powerful test for location or scale alternatives from any specified distribution. Another important application is in the development of asymptotically fully efficient tests for multiple or stratified 2×2 tables under various scales or measures of association. This application is addressed in Chapter 4. The following section presents an illustration of these concepts. Additional problems involving evaluation of the power and sample size for specific tests, and the evaluation of the Pitman efficiency, are presented in future chapters.

3.6.4 Stratified Versus Unstratified Analysis of Risk Differences

To illustrate these concepts, consider a study where we wish to assess the differences in proportions between two treatment or exposure groups where we may also be interested in adjusting for the differences in risk between subgroups or strata. Such analyses are the subject of Chapter 4. Here, we consider the simplest case of two strata, such as men and women, within which the underlying probabilities or risks of the outcome may differ. We then wish to test the null hypothesis of no difference within both strata $H_0: \pi_{1j} = \pi_{2j} = \pi_j$ for $j = 1, 2$, versus the alternative $H_1: \theta = \pi_{1j} - \pi_{2j} \neq 0$ of a constant difference across strata, where the probability of the positive outcome in the control group differs between strata, $\pi_{21} \neq \pi_{22}$. Under this model we then wish to evaluate the *ARE* of an unstratified Z -test based on the pooled 2×2 table versus a stratified-adjusted test.

Under this model, the exposed or treated group probability in the j th stratum is $\pi_{1j} = \theta + \pi_{2j}$ for $j = 1, 2$. Let ζ_j denote the expected sample fraction in the j th stratum, $E(N_j/N) = \zeta_j$, where N_j is the total sample size in the j th stratum ($j = 1, 2$) and N is the overall total sample size. Within the j th stratum we assume that there are equal sample sizes within each group by design such that $n_{ij} = N_j/2$ and $E(n_{ij}) = \zeta_j N/2$ ($i = 1, 2$; $j = 1, 2$). The underlying model can then be

summarized as follows:

Stratum	Stratum Fraction	Probability + Exposed	Probability + Control
1	ζ_1	$\theta + \pi_{21}$	π_{21}
2	ζ_2	$\theta + \pi_{22}$	π_{22}

where $\zeta_1 + \zeta_2 = 1$.

Within each stratum let p_{ij} refer to the proportions positive in the i th group in the j th stratum, where $E(p_{ij}) = \pi_{ij}$ ($i = 1, 2$; $j = 1, 2$). Then the unstratified Z -test simply employs (2.83) using the $\{p_{i\bullet}\}$ and $\{n_{i\bullet}\}$ from the pooled 2×2 table where $n_{i\bullet} = N/2$ by design and

$$p_{i\bullet} = \frac{n_{i1}p_{i1} + n_{i2}p_{i2}}{n_{i1} + n_{i2}} = \frac{N_1p_{i1} + N_2p_{i2}}{N}, \quad (3.90)$$

since $n_{ij} = N_j/2$ and $N_1 + N_2 = N$. Thus, the test statistic in the marginal unadjusted analysis is

$$T_\bullet = p_{1\bullet} - p_{2\bullet} = \sum_{j=1}^2 \frac{N_j}{N} (p_{1j} - p_{2j}). \quad (3.91)$$

Under this model, the underlying parameters are the probabilities within each stratum, the $\{\pi_{ij}\}$. Thus, $E(p_{i\bullet})$ is a function of the $\{\pi_{ij}\}$ and the sampling fractions of the two strata:

$$\begin{aligned} E(p_{2\bullet}) &= \sum_{j=1}^2 \zeta_j \pi_{2j} = \pi_{2\bullet} \\ E(p_{1\bullet}) &= \sum_{j=1}^2 \zeta_j (\theta + \pi_{2j}) = \pi_{1\bullet} = \theta + \pi_{2\bullet}. \end{aligned} \quad (3.92)$$

Thus,

$$E(T_\bullet) = E(p_{1\bullet}) - E(p_{2\bullet}) = \theta \quad (3.93)$$

and the statistic is asymptotically unbiased.

The variance of this pooled statistic is $V(T_\bullet) = V(p_{1\bullet}) + V(p_{2\bullet})$. Under H_0 : $\pi_{1j} = \pi_{2j} = \pi_j$, then

$$V(p_{ij}|H_0) = \frac{\pi_j(1 - \pi_j)}{N\zeta_j/2}. \quad (3.94)$$

Thus,

$$V(p_{i\bullet}|H_0) = \sum_{j=1}^2 \frac{\zeta_j^2 \pi_j(1 - \pi_j)}{N\zeta_j/2} \quad (3.95)$$

and

$$V(T_\bullet|H_0) = \sum_{j=1}^2 \zeta_j^2 \pi_j(1 - \pi_j) \left(\frac{4}{N\zeta_j} \right) = \sum_{j=1}^2 \zeta_j^2 \sigma_{0j}^2, \quad (3.96)$$

where

$$\sigma_{0j}^2 = \pi_j(1 - \pi_j) \left(\frac{4}{N\zeta_j} \right) \quad (3.97)$$

is the variance of the difference within the j th stratum under the null hypothesis with variance component $\phi_{0j}^2 = 4\pi_j(1 - \pi_j)/\zeta_j$.

Now consider the stratified-adjusted test where the test statistic is an optimally weighted average of the differences within each of the two strata of the form

$$T = \sum_{j=1}^2 w_j(p_{1j} - p_{2j}), \quad (3.98)$$

where $w_1 + w_2 = 1$. First, the weights must be determined that provide maximum efficiency (power) under the specified model that there is a constant difference within the two strata. Then, given the optimal weights, we obtain the asymptotic efficiency of the test and the asymptotic relative efficiency versus that of another test, such as the pooled unstratified test above.

Since $E(p_{1j} - p_{2j}) = \theta$ within each stratum ($j = 1, 2$), then asymptotically $E(T) = \theta$ for any set of weights that sum to 1. The variance of the statistic, however, depends explicitly on the chosen weights through

$$V(T|H_0) = \sum_{j=1}^2 w_j^2 \pi_j(1 - \pi_j) \left(\frac{4}{N\zeta_j} \right) = \sum_{j=1}^2 w_j^2 \sigma_{0j}^2. \quad (3.99)$$

Since the statistic provides a consistent estimate of θ , then the asymptotic efficiency of the test is

$$Eff(T) = \frac{1}{V(T|H_0)} = \frac{1}{w_1^2 \sigma_{01}^2 + w_2^2 \sigma_{02}^2}. \quad (3.100)$$

To obtain the test with greatest power asymptotically, we desire that value of w_1 (and w_2) for which the asymptotic efficiency is maximized. Using the calculus of maxima/minima, it is readily shown (see Problem 3.6) that the optimal weights are defined as

$$w_j = \frac{\sigma_{0j}^{-2}}{\sigma_{01}^{-2} + \sigma_{02}^{-2}}, \quad (3.101)$$

which are inversely proportional to the variance of the difference within the j th stratum. With these weights, it follows that

$$V(T|H_0) = \frac{1}{\sigma_{01}^{-2} + \sigma_{02}^{-2}}, \quad (3.102)$$

which is the minimum variance for any linear combination over the two strata. These results generalize to the case of more than two strata. Such weights that are *inversely proportional to the variance* play a central role in optimally weighted estimates and tests.

We can now assess the asymptotic relative efficiency of the stratified-adjusted test with optimal weights versus the pooled test. From (3.92)–(3.93), the weights in the pooled test asymptotically are simply the sample fractions $\{\zeta_j\}$. Since both test

statistics are based on a consistent estimator of θ , then asymptotically $dE(T_*)/d\theta = dE(T)/d\theta = 1$ when evaluated at any value of θ . Then

$$\begin{aligned} ARE(T_*, T) &= \left[\frac{V(T_*)|\theta|}{V(T|\theta)} \right]_{|\theta=0} = \frac{w_1^2 \sigma_{01}^2 + w_2^2 \sigma_{02}^2}{\zeta_1^2 \sigma_{01}^2 + \zeta_2^2 \sigma_{02}^2} \\ &= \frac{[\sigma_{01}^{-2} + \sigma_{02}^{-2}]^{-1}}{\zeta_1^2 \sigma_{01}^2 + \zeta_2^2 \sigma_{02}^2} = \frac{[\phi_{01}^{-2} + \phi_{02}^{-2}]^{-1}}{\zeta_1^2 \phi_{01}^2 + \zeta_2^2 \phi_{02}^2}. \end{aligned} \quad (3.103)$$

Because the total sample size N cancels from each variance expression, the ARE is not a function of the sample size.

By construction, the optimal weights minimize the variance of a linear combination over strata, so that $ARE(T_*, T) \leq 1$ and the stratified-adjusted test will always provide greater power when there is a constant difference over strata. However, when the difference between groups is not constant over strata and the model assumptions do not apply, then the stratified-adjusted test may be less powerful than the pooled test. In Chapter 4 we consider these issues in greater detail. There we generalize the analysis to allow for any number of strata, and we consider hypotheses that there is a constant difference on other scales, such as a constant relative risk or constant odds ratio over multiple 2×2 tables.

Example 3.10 Two Strata

Consider the case of two strata with the following parameters:

Stratum (j)	ζ_j	π_{1j}	π_{2j}	ϕ_{0j}^2	w_j
1	0.3	0.35	0.2	2.6583	0.34898
2	0.7	0.55	0.4	1.4250	0.65102

where ϕ_{0j}^2 is the variance component from (3.97) that does not depend on the total sample size. In this case, $ARE(T_*, T) = 0.98955$, indicating a slight loss of efficiency when using the pooled test versus the stratified-adjusted test. Note that the optimal weights are close to the fractions within each stratum, $w_j \doteq \zeta_j$, so that there is similar efficiency of the two tests.

In Section 4.4.3, however, we show that the marginal unadjusted statistic $T_* = p_{1*} - p_{2*}$ is biased when there are differences in the sample fractions within each treatment group among strata; that is, when $n_{i1}/N_1 \neq n_{i2}/N_2$ for the i th treatment group with the two strata. In this case, the marginal unadjusted analysis would use a bias-corrected statistic, and the relative efficiency of the stratified-adjusted test would be much greater.

3.7 PROBLEMS

3.1 Consider a one-sided Z -test against a left-tailed alternative hypothesis. Derive the basic equation (3.18) that describes the power for this test.

3.2 Consider the Pearson contingency chi-square test for a 2×2 table.

- 3.2.1. For all four cells of the table, show that $[\pi_{ij} - \pi_{0ij}]^2 = \delta^2$ for $\delta \neq 0$.
 3.2.2. Show that the noncentral factor for the X^2 test can be expressed as

$$E(X^2) = N\delta^2 \sum_{i=1}^2 \sum_{j=1}^2 \frac{1}{\pi_{0ij}}. \quad (3.104)$$

3.2.3. Show that this equation is equivalent to $\psi^2 = N\tau^2$, where τ is provided by the simplification in (3.39).

3.3 Consider a study to assess the difference between two groups in the incidence of improvement (healing) among those treated with a drug versus placebo, where prior studies suggest that the placebo control group probability is about $\pi_2 = 0.20$. The investigators then wish to detect an minimal increase in the probability of healing with drug treatment on the order of 0.10, that is, a probability of $\pi_1 = 0.30$. Perform the following calculations needed to design the trial.

3.3.1. For a two-sided test at the $\alpha = 0.05$ -level with $Z_{0.975} = 1.96$, what total sample size N would be needed (with equal-sized groups) to provide power $1 - \beta = 0.90$ ($Z_{0.90} = 1.282$) to detect this difference?

3.3.2. Now suppose that a total sample size of only $N = 400$ is feasible. With $\pi_2 = 0.2$, what level of power is there to detect

1. A difference of 0.10?
2. A relative risk of 1.5?
3. An odds ratio of 2.0?

When doing this, note that as π_1 changes, so does π .

3.3.3. Now suppose that the control healing rate is actually higher than originally expected, say $\pi_2 = 0.30$ rather than the initial projection of 0.20. For a total sample size of $N = 400$ with equal-sized groups, recompute the power to detect a difference of 0.10, a relative risk of 1.5, and an odds ratio of 2.0.

3.3.4. Suppose that the new treatment is expensive to administer. To reduce costs the sponsor requires that only 1/3 of the total N be assigned to the experimental treatment ($\xi_1 = 1/3$). Recompute the power for the conditions in Problems 3.3.2 and 3.3.3. What effect does the unbalanced design have on the power of the test?

3.4 Consider the case of the large sample Z -test for the difference between the means of two populations based on the difference between two sample means \bar{x}_1 and \bar{x}_2 that are based on samples drawn from some distribution with equal variances φ^2 in each population. Then, asymptotically,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\varphi \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \stackrel{d}{\approx} \mathcal{N}(0, 1) \quad (3.105)$$

under the null hypothesis $H_0: E(\bar{x}_1) = E(\bar{x}_2)$. Let $E(\bar{x}_i) = v_i$ for $i = 1, 2$.

- 3.4.1. Show the asymptotic distribution of Z under $H_1: v_1 \neq v_2$.

3.4.2. Derive the equation to compute the sample size N required to detect a difference $\mu_1 = v_1 - v_2 \neq 0$ expressed as

$$N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \varphi^2 \left(\frac{1}{\xi_1} + \frac{1}{\xi_2} \right)}{(v_1 - v_2)^2}. \quad (3.106)$$

with sample fractions $\{\xi_i\}$.

3.4.3. Derive the equation to compute the power of the test expressed as

$$Z_{1-\beta} = \frac{\sqrt{N} (v_1 - v_2)}{\varphi \sqrt{\frac{1}{\xi_1} + \frac{1}{\xi_2}}} - Z_{1-\alpha} = \frac{v_1 - v_2}{\varphi \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - Z_{1-\alpha}. \quad (3.107)$$

3.4.4. What sample size is required to detect a difference $\mu_1 = 0.20$ with power = 0.9 where $\varphi^2 = 1$ with two equal-sized groups using a two-sided test at level $\alpha = 0.05$, two-sided?

3.4.5. What power would be provided to detect this difference with $N = 120$?

3.5 Consider the large sample Z -test of the difference between the rate or intensity parameters for two populations, each with a homogeneous Poisson distribution with rate parameters λ_1 and λ_2 , where there is an equal period of exposure per observation (see Section 8.1.2). For a sample of n_i observations in the i th group ($i = 1, 2$), the sample estimate of the rate is $\hat{\lambda}_i = d_i/n_i$, where d_i is the number of events observed among the n_i observations in the i th group. Under H_0 : $\lambda_1 = \lambda_2 = \lambda$, the sample estimate of the assumed common rate is $\hat{\lambda} = (d_1 + d_2)/(n_1 + n_2)$.

3.5.1. Within the i th group, from the normal approximation to the Poisson, $d_i \stackrel{d}{\approx} \mathcal{N}(n_i \lambda_i, n_i \lambda_i)$, show that under H_0 the large sample Z -test is

$$Z = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\hat{\lambda} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim \mathcal{N}(0, 1). \quad (3.108)$$

3.5.2. Show the asymptotic distribution of $\hat{\lambda}_1 - \hat{\lambda}_2$ under H_1 : $\lambda_1 \neq \lambda_2$.

3.5.3. Derive the equation to compute the sample size N required to detect a difference $\mu_1 = \lambda_1 - \lambda_2 \neq 0$, expressed as

$$N = \left(Z_{1-\alpha} \sqrt{\lambda \left(\frac{1}{\xi_1} + \frac{1}{\xi_2} \right)} + Z_{1-\beta} \sqrt{\frac{\lambda_1}{\xi_1} + \frac{\lambda_2}{\xi_2}} \right)^2 / (\lambda_1 - \lambda_2)^2. \quad (3.109)$$

3.5.4. Derive the equation to compute the power of the test expressed as

$$Z_{1-\beta} = \frac{\sqrt{N} |\lambda_1 - \lambda_2| - Z_{1-\alpha} \sqrt{\lambda \left(\frac{1}{\xi_1} + \frac{1}{\xi_2} \right)}}{\sqrt{\frac{\lambda_1}{\xi_1} + \frac{\lambda_2}{\xi_2}}}. \quad (3.110)$$

3.5.5. What sample size is required to detect a difference $\lambda_1 = 0.20$ versus $\lambda_2 = 0.35$ with 90% power in a study with two equal-sized groups using a two-sided test at level $\alpha = 0.05$?

3.5.6. What power would be provided to detect this difference with $N = 200$?

3.6 Consider an equivalence design specified in terms of a relative risk. For given probabilities (π_1, π_2) , a symmetric two-sided margin on the $\log(RR)$ is specified as $(-\theta_E, \theta_E)$, where $\theta_E = \ln(\pi_1/\pi_2)$. Then equivalence is declared if

$$\ln(p_1/p_2) \pm Z_{1-\alpha/2}\sigma_0 \in (-\theta_E, \theta_E), \quad (3.111)$$

where σ_0^2 is as shown in (2.43). For sample fractions (ξ_1, ξ_2) , then

$$\sigma_0^2 = \frac{1 - \pi}{\pi N \xi_1 \xi_2} = \phi_0^2/N. \quad (3.112)$$

3.6.1. Show that the margin satisfies

$$\theta_E = (Z_{1-\alpha/2} + Z_{1-\beta/2})\sigma_0 \quad (3.113)$$

and that the required sample size is obtained as

$$N = \left[\frac{(Z_{1-\alpha/2} + Z_{1-\beta/2})\phi_0}{\theta_E} \right]^2. \quad (3.114)$$

3.6.2. Assume that $\pi = 0.6$ and that the margin specifies an upper $RR = 1.1$ so that $\theta_E = \ln(1.1)$. For $\alpha = 0.05$, $1 - \beta = 0.90$, what sample size is required?

3.6.3. Assume that a noninferiority design is to be implemented using a bound on the relative risk as above. Show that the required sample size is provided by (3.114) but using the one-sided quantiles $Z_{1-\alpha}$ (for a one-sided limit) and $Z_{1-\beta}$.

3.6.4. Again assume that $\pi = 0.6$ and that $\theta_E = \ln(1.1)$. For $\alpha = 0.05$, $1 - \beta = 0.90$, what sample size is required?

3.7 Consider the *ARE* of the stratified versus marginal analysis in Section 3.5.4.

3.7.1. Show that for the stratified-adjusted test, the efficiency of the test is maximized and the variance in (3.100) is minimized using the optimal weights in (3.101).

3.7.2. Then show that $V(T|H_0)$ equals the expression in (3.102).

3.7.3. Now consider three strata where the variance of the stratified-adjusted statistic is as in (3.99), with summation over the three strata. Then we desire the weights w_1 and w_2 that minimize the function

$$V(T|H_0) = w_1\sigma_{01}^2 + w_2\sigma_{02}^2 + (1 - w_1 - w_2)\sigma_{03}^2. \quad (3.115)$$

Differentiating with respect to w_1 and w_2 yields two simultaneous equations. Solve these and then simplify to show that the optimal weights are

$$w_j = \frac{\sigma_{0j}^{-2}}{\sigma_{01}^{-2} + \sigma_{02}^{-2} + \sigma_{03}^{-2}} \quad (3.116)$$

for $j = 1, 2, 3$.

Stratified-Adjusted Analysis for Independent Groups

4.1 INTRODUCTION

In many studies it is important to account for or adjust for the potential influence of other covariates on the observed association between the treatment or exposure groups and the response. This observed association may be biased when there is an imbalance in the distributions of an important covariate between the two groups. One approach to adjust for such imbalances is to conduct a stratified analysis, stratifying on the other covariates of importance. Another, introduced in later chapters, is to employ an appropriate regression model. When the regression model is equivalent to the model adopted in the stratified analysis, then the two approaches are equivalent, at least asymptotically.

In a stratified analysis the original samples of n_1 and n_2 observations from two groups are divided into strata, each an independent subdivision of the study, for example, males and females, and for each a separate 2×2 table is constructed. The stratified-adjusted analysis then aggregates the measures of association between group and response across all strata to provide a stratified-adjusted measure of association. The analysis also provides an aggregate overall stratified-adjusted test of significance. The stratified-adjusted analysis is also called an analysis of *partial association* because it assesses the influence of the treatment or exposure group on the response, after allowing, or adjusting, for the association between the stratifying covariate with both group membership and with the response.

In the Chapter 3 we explored a stratified test of significance of risk differences over two strata to illustrate the concept of the asymptotic relative efficiency of two possible tests for the same hypothesis. We now extend this concept to include estimators of an assumed common or average parameter over strata and asymptotically efficient tests of significance. In addition to the risk difference, we consider the log relative risk

and the log odds ratio. We do so first from the perspective of a fixed effects model and then generalize these developments to a random effects model.

A stratified analysis may be performed for various reasons. The most common is to adjust for the potential influence of an important covariate that may be associated with group membership. In observational studies, it is possible that such a covariate may explain some of the observed association between group and response, or the lack thereof. This can occur when there is an imbalance between groups with respect to the distribution of a covariate, such as when there is a larger fraction of males in one group than in the other. However, in a stratified analysis this imbalance can be accounted for by comparing the groups separately within each stratum, such as separately among males and separately among females. In this manner the influence of the covariate on the observed association, if any, between group and response is accounted for or adjusted for.

Another common application of a stratified analysis is a *meta-analysis*. This refers to the combination of the results of separate studies or substudies, each constituting a separate stratum in the analysis. The analysis then provides a stratified-adjusted combined estimate of an overall group effect and assesses its statistical significance. For example, in a multicenter clinical trial, it may be desirable to conduct a combined assessment of the overall treatment effect adjusting for center-to-center differences in the nature of the patient populations or other center-specific differences. In this case, the results within each clinical center form an independent stratum and the adjusted analysis combines the results over strata. Similarly, a meta-analysis of a specific treatment effect may be obtained by conducting a stratified-adjusted analysis in which the results of several published studies of the same treatment versus control are combined. For example, the Early Breast Cancer Trialist's Collaborative Group (1998) presented a meta-analysis of 37,000 women from 55 studies of the effects of adjuvant tamoxifen on the risk of recurrence of breast cancer following surgical treatment. The majority of these studies had produced inconclusive (not statistically significant) results. However, combining all of these studies into a single analysis greatly increased the power to detect an important therapeutic effect. Among all women, the risk of cancer recurrence was reduced by 26%, and among those treated for three or more years it was reduced by 42%, each $p \leq 0.0001$.

The most commonly used and well-known method for conducting a stratified analysis of multiple independent 2×2 tables is the procedure of Mantel and Haenszel (1959). The Mantel-Haenszel procedure yields an aggregate or combined test of partial association that is optimal under the alternative hypothesis of a common odds ratio. An asymptotically equivalent test for a common odds ratio is the test of Cochran (1954a). Cochran's test is also a member of a family of tests described by Radhakrishna (1965), which also includes tests of a common relative risk or a common risk difference, among others. Mantel and Haenszel (1959) also describe an estimate of the assumed common odds ratio. Other estimators are the maximum likelihood estimates, and a family of efficient estimates described by Gart (1971) that can be derived using weighted least squares. These and other related procedures are the subject of this chapter.

First we consider the analyses within strata and the unadjusted marginal analysis. We then consider the asymptotically equivalent Mantel-Haenszel and Cochran stratified-adjusted tests of significance.

4.2 MANTEL-HAENSZEL TEST AND COCHRAN'S TEST

4.2.1 Conditional Within-Strata Analysis

The analysis begins with a conditional within-strata analysis in which a separate 2×2 table is constructed from the observations within each strata. Let K refer to the total number of strata, indexed by $k = 1, \dots, K$. The strata may be defined by the categories of a single covariate, for example, gender, or by the intersection of the categories of two or more covariates considered jointly, such as by four categories defined by gender and the presence versus absence of a family history of diabetes. Conditionally within the k th strata, the observed frequencies in the 2×2 table and the corresponding probabilities of a positive response with each group are denoted as

(4.1)

		kth Stratum		Probabilities	
		Group		Group	
Response	+	1	2	1	2
		a_k	b_k	π_{1k}	π_{2k}
	-	c_k	d_k	$1 - \pi_{1k}$	$1 - \pi_{2k}$
		n_{1k}	n_{2k}	1	1
		N_k			

Within the k th stratum, any of the measures of group-response association described previously may be computed: the risk difference $\widehat{RD}_k = p_{1k} - p_{2k}$, relative risk $\widehat{RR}_k = p_{1k}/p_{2k} = (a_k/n_{1k})/(b_k/n_{2k})$, and the odds ratio $\widehat{OR}_k = p_{1k}/(1 - p_{1k})/[p_{2k}/(1 - p_{2k})] = a_k d_k / b_k c_k$. One could also conduct a test of association using the contingency chi-square test (Cochran's test), Mantel-Haenszel test, or an exact test, separately for each of the K tables. However, this leads to the problem of multiple tests of significance, with pursuant inflation of the Type I error probability for the set of tests. Thus, it is preferable to obtain a single overall test of significance for the adjusted association between group and response, and a single overall estimate of the degree of association.

4.2.2 Marginal Unadjusted Analysis

In Chapter 2 we considered the case of a single 2×2 table. This is also called a marginal or unadjusted analysis because it is based on the pooled data for all strata combined. Using the " \cdot " notation to refer to summation over the K strata, the

observed frequencies and the underlying probabilities for the marginal 2×2 table are

<u>Frequencies</u>		<u>Probabilities</u>	
<i>Group</i>		<i>Group</i>	
	1	2	
+	a_{\bullet}	b_{\bullet}	$m_{1\bullet}$
-	c_{\bullet}	d_{\bullet}	$m_{2\bullet}$
	$n_{1\bullet}$	$n_{2\bullet}$	N_{\bullet}

<u>Frequencies</u>		<u>Probabilities</u>	
<i>Group</i>		<i>Group</i>	
	1	2	
+	$\pi_{1\bullet}$	$\pi_{2\bullet}$	
-	$1 - \pi_{1\bullet}$	$1 - \pi_{2\bullet}$	
	1	1	

This provides the basis for the unadjusted conditional Mantel-Haenszel test statistic $X_{c\bullet}^2$ in (2.86) and the unconditional Cochran test statistic $X_{u\bullet}^2$ in (2.93), the "•" designating the marginal test. The measures of association computed for this table are likewise designated as \widehat{RD}_{\bullet} , \widehat{RR}_{\bullet} , and \widehat{OR}_{\bullet} . These tests and estimates, however, ignore the possible influence of the stratifying covariate.

Example 4.1 *Clinical Trial in Duodenal Ulcers*

Blum (1982) describes a hypothetical clinical trial of the effectiveness of a new drug versus placebo. Through randomization, the study is expected to provide an unbiased assessment of the difference between treatments, so that the marginal unadjusted analysis is expected to be unbiased. Even in this case, however, it is sometimes instructive to assess the treatment effect adjusting for an important covariate that may have a strong influence on the risk of the outcome or the effectiveness of the treatment. In this example we assess the effectiveness of a new drug for the treatment of duodenal ulcers, where the drug is expected to promote healing by retarding the excretion of gastric juices that lead to ulceration of the duodenum.

Ulcers typically have three classes of etiology. *Acid-dependent* ulcers are principally caused by excessive gastric secretion, and it is expected that these ulcers will be highly responsive to a treatment that reduces such secretion. *Drug-dependent* ulcers are usually formed by excessive use of drugs, such as aspirin, that may irritate the lining of the duodenum. Since gastric secretion plays a minor role in drug-induced ulcer formation, it is expected that these ulcers will be resistant to treatment. Ulcers of *intermediate origin* are those where it is difficult to determine whether the principal cause is a result of acid secretion or excessive use of a drug irritant.

Initially, 100 patients were assigned to each treatment (drug vs. placebo). When stratified by ulcer type, the following 2×2 tables for drug (D) versus placebo (P) and healing (+) versus not (-) are formed.

$$\begin{array}{l}
 \begin{array}{c} 1. \text{ Acid-Dependent} \\ \begin{array}{cc} D & P \\ \hline 16 & 20 \\ 26 & 27 \end{array} \end{array} \quad \begin{array}{c} 2. \text{ Drug-Dependent} \\ \begin{array}{cc} D & P \\ \hline 9 & 4 \\ 3 & 5 \end{array} \end{array} \quad \begin{array}{c} 3. \text{ Intermediate} \\ \begin{array}{cc} D & P \\ \hline 28 & 16 \\ 18 & 28 \end{array} \end{array} \\
 + \quad \quad \quad + \quad \quad \quad + \quad \quad \quad (4.3) \\
 \begin{array}{cc} 36 & \\ \hline 53 & \end{array} \quad \begin{array}{cc} 13 & \\ \hline 8 & \end{array} \quad \begin{array}{cc} 44 & \\ \hline 46 & \end{array} \\
 \begin{array}{cc} 42 & 47 \\ \hline 89 & \end{array} \quad \begin{array}{cc} 12 & 9 \\ \hline 21 & \end{array} \quad \begin{array}{cc} 46 & 44 \\ \hline 90 & \end{array}
 \end{array}$$

(reproduced with permission). The unadjusted analysis is based on the marginal 2×2 table, obtained as the sum of the stratum-specific tables, to yield

		Marginal	
		D	P
+	53	40	93
	47	60	107
100	100	200	

Within each stratum, and in the marginal unadjusted analysis among all strata combined, the proportions that healed in each group are

Group Proportion	Stratum			Marginal Unadjusted
	1	2	3	
Drug (p_1)	0.381	0.750	0.609	0.530
Placebo (p_2)	0.426	0.444	0.364	0.400

The estimates of the three principal summary measures and the 95% confidence limits are presented in Table 4.1. For the relative risk and odds ratios, the asymmetric confidence limits are presented. The Cochran and Mantel-Haenszel test statistics within each stratum and marginally (unadjusted) are

Test	Stratum			Marginal Unadjusted
	1	2	3	
Mantel-Haenszel X_c^2	0.181	1.939	5.345	3.380
$p \leq$	0.671	0.164	0.021	0.067
Cochran X_u^2	0.183	2.036	5.405	3.397
$p \leq$	0.669	0.154	0.021	0.066

Among those treated with the drug, the highest healing rates (proportions) were observed in the drug-dependent (2) and intermediate (3) ulcer strata. The beneficial effect of drug treatment was greatest among those with drug-dependent ulcers, whether measured as the risk difference, relative risk, or odds ratio. Although the tests of significance within strata are presented, these are not usually employed because of the problem of multiple tests of significance and the pursuant increase in the Type I error probability.

Marginally, the differences between groups is not statistically significant. This test, however, ignores the imbalances between groups in the numbers of subjects within each of the three ulcer-type strata. One might ask whether the nature or significance of the treatment group effect is altered in any way after adjusting for these imbalances within strata.

4.2.3 Mantel-Haenszel Test

Mantel and Haenszel (1959) advanced the following test for a common odds ratio among the set of K tables that adjusts for the influence of the stratifying covariate(s).

Table 4.1 Measures of association within each stratum and in the marginal unadjusted analysis.

Measure	Stratum			Marginal
	1	2	3	
Risk difference (\widehat{RD})	-0.045	0.306	0.245	0.130
$\widehat{V}(\widehat{RD})$	0.011	0.043	0.010	0.0049
95% C.I. for RD	-0.25, 0.16	-0.10, 0.71	0.04, 0.45	-0.007, 0.27
Relative risk (\widehat{RR})	0.895	1.688	1.674	1.325
log relative risk	-0.111	0.523	0.515	0.281
$\widehat{V}[\log \widehat{RR}]$	0.067	0.167	0.054	0.0239
95% C.I. for RR	0.54, 1.49	0.76, 3.76	1.06, 2.64	0.98, 1.79
Odds ratio (\widehat{OR})	0.831	3.750	2.722	1.691
log odds ratio	-0.185	1.322	1.001	0.526
$\widehat{V}[\log \widehat{OR}]$	0.188	0.894	0.189	0.0818
95% C.I. for OR	0.36, 1.94	0.59, 23.9	1.16, 6.39	0.97, 2.96

This provides a test of the global null hypothesis $H_0: \pi_{1k} = \pi_{2k}$ ($OR_k = 1$) for all k versus the alternative that the probabilities within strata differ such that there is a common odds ratio $\neq 1$. That is, $H_1: OR_k = OR \neq 1$ for all $k = 1, \dots, K$. As shown in Section 2.6.3, from the conditional hypergeometric likelihood for a single 2×2 table under the null hypothesis within the k th stratum, the expected frequency for the index cell, $E(a_k)$, is simply

$$E_k = E(a_k) = n_{1k}m_{1k}/N_k \quad (4.4)$$

and central hypergeometric variance of a_k under H_0 is

$$V_{ck} = \frac{m_{1k}m_{2k}n_{1k}n_{2k}}{N_k^2(N_k - 1)}. \quad (4.5)$$

Within the k th stratum, these provide a within-stratum test. However, we are principally interested in the aggregate stratified-adjusted test. The Mantel-Haenszel test and its asymptotic distribution are obtained as follows.

Within the k th stratum, under H_0 asymptotically for large N_k and for fixed K ,

$$a_k - E_k \xrightarrow{d} \mathcal{N}[0, V_{ck}]. \quad (4.6)$$

Since the strata are independent,

$$\sum_k (a_k - E_k) \xrightarrow{d} \mathcal{N}\left[0, \sum_k V_{ck}\right]. \quad (4.7)$$

Therefore, the stratified-adjusted Mantel-Haenszel test, conditional on all margins fixed, is

$$X_{C(MH)}^2 = \frac{\left[\sum_k (a_k - E_k) \right]^2}{\sum_k V_{ck}} = \frac{\left[\sum_k (a_k - n_{1k}m_{1k}/N_k) \right]^2}{\sum_k \left(\frac{m_{1k}m_{2k}n_{1k}n_{2k}}{N_k^2(N_k - 1)} \right)} = \frac{\left[a_+ - E_+ \right]^2}{V_{c+}}, \quad (4.8)$$

where $a_+ = \sum_{k=1}^K a_k$, $E_+ = \sum_k E_k$ and $V_{c+} = \sum_k V_{ck}$. Since $a_+ - E_+$ is the sum of asymptotically normally distributed stratum-specific variates, then asymptotically $(a_+ - E_+) \xrightarrow{d} \mathcal{N}[0, V_{c+}]$ and asymptotically $X_{C(MH)}^2 \xrightarrow{d} \chi^2$ on 1 df. Note that while $a_+ = a_*$ is also the index frequency in the marginal 2×2 table in (4.2), the other components E_+ and V_{c+} are not based on the marginal table at all.

The asymptotic distribution can also be demonstrated for the case where the sample size within each stratum is small but the number of strata increases indefinitely (Breslow, 1981).

4.2.4 Cochran's Test

An asymptotically equivalent test was also suggested by Cochran (1954a) using a linear combination of the differences between the proportions $p_{1k} - p_{2k}$ over the K strata. The representation of this and other tests in this form is described subsequently. Algebraically, Cochran's stratified-adjusted test can also be expressed in terms of $a_+ - E_+$.

Starting from the unconditional product binomial likelihood, as shown in Section 2.6.4 for a single 2×2 table, then under the null hypothesis, $E(a_k) = n_{1k}\pi_k$ within the k th stratum, that can be estimated consistently as

$$\hat{E}_k = n_{1k}m_{1k}/N_k. \quad (4.9)$$

As shown in (2.91), the unconditional variance of a_k is

$$V_{uk} = \frac{n_{1k}n_{2k}\pi_k(1 - \pi_k)}{N_k}, \quad (4.10)$$

which can be consistently estimated as

$$\hat{V}_{uk} = \frac{m_{1k}m_{2k}n_{1k}n_{2k}}{N_k^3}. \quad (4.11)$$

Therefore, when aggregated over strata, the stratified-adjusted unconditional test is provided by

$$X_{U(C)}^2 = \frac{\left[\sum_k \left(a_k - \hat{E}_k \right) \right]^2}{\sum_k \hat{V}_{uk}} = \frac{\left[\sum_k (a_k - n_{1k}m_{1k}/N_k) \right]^2}{\sum_k (m_{1k}m_{2k}n_{1k}n_{2k}/N_k^3)} = \frac{\left[a_+ - \hat{E}_+ \right]^2}{\hat{V}_{u+}}. \quad (4.12)$$

Analogously to (4.6) and (4.7), asymptotically $(a_k - E_k) \sim N[0, V_{uk}^2]$ within the k th stratum and $(a_+ - E_+) \stackrel{d}{\approx} \mathcal{N}[0, V_{u+}^2]$. Extending the approach used in Section (2.6.4), asymptotically $\widehat{E}_+ \xrightarrow{p} E_+$ and $\widehat{V}_{u+} \xrightarrow{p} V_{u+}$, and from Slutsky's theorem (Section A.4.3), $X_{U(C)}^2 \stackrel{d}{\approx} \chi^2$ on 1 df .

Since the only difference between the conditional Mantel-Haenszel test $X_{C(MH)}^2$ and the unconditional Cochran test $X_{U(C)}^2$ is in the denominators, and since $\widehat{V}_{uk} = V_{ck}(N_k - 1)/N_k$, then the two tests are asymptotically equivalent. Thus, the two are often referred to interchangeably as the *Cochran-Mantel-Haenszel test*.

Formally, as will be shown in Section 4.7, these tests address the following null hypothesis:

$$H_0: OR = 1 \text{ assuming } E(\widehat{OR}_k) = OR_k = OR, \forall k \quad (4.13)$$

versus the alternative

$$H_1: OR \neq 1 \text{ assuming } E(\widehat{OR}_k) = OR_k = OR, \forall k.$$

Thus, these tests are described explicitly in terms of odds ratios. This also implies that different test statistics could be used to detect an adjusted difference in relative risks or risk differences. Such tests are also described in Section 4.7.

Example 4.2 Clinical Trial in Duodenal Ulcers (continued)

For the above example, the Mantel-Haenszel test statistic is $X_{C(MH)}^2 = 3.0045$ with $p \leq 0.083$. Cochran's test is $X_{U(C)}^2 = 3.0501$ with $p \leq 0.081$. Compared to the marginal unadjusted analysis, the stratified-adjusted analysis yields a slightly less significant test statistic value, $p \leq 0.083$ versus $p \leq 0.067$ (using the Mantel-Haenszel tests). These results suggest that some of the association between treatment and response is now accounted for through the stratification adjustment. Stratified-adjusted estimates of the degree of association are then required to further describe the nature of the stratification adjustment.

4.3 STRATIFIED-ADJUSTED ESTIMATORS

4.3.1 Mantel-Haenszel Estimates

Mantel and Haenszel (1959) presented a "heuristic" estimate of the assumed common odds ratio, and also of the assumed relative risks when a common relative risk is assumed to exist rather than a common odds ratio. There was no formal proof of the asymptotic properties of these estimators, nor were they derived so as to provide any particular statistical properties. Rather, Mantel and Haenszel simply stated that these estimators seemed to work well compared to other possible estimators. In fact, it was not until recently that the asymptotic variances of the estimates were actually derived.

The Mantel-Haenszel estimate of the common odds ratio is

$$\widehat{OR}_{MH} = \frac{\sum_k a_k d_k / N_k}{\sum_k b_k c_k / N_k}, \quad (4.14)$$

which can be expressed as a weighted combination of the stratum-specific estimates of the form

$$\widehat{OR}_{MH} = \sum_k \widehat{v}_k (\widehat{OR}_k) \quad (4.15)$$

with weights

$$\widehat{v}_k = \frac{b_k c_k / N_k}{\sum_\ell b_\ell c_\ell / N_\ell} \quad (4.16)$$

that sum to unity.

Likewise, when a constant relative risk is assumed, the Mantel-Haenszel estimate of the common relative risk is

$$\widehat{RR}_{MH} = \frac{\sum_k a_k n_{2k} / N_k}{\sum_k b_k n_{1k} / N_k}. \quad (4.17)$$

This estimate can also be expressed as a weighted average of the stratum-specific relative risks:

$$\widehat{RR}_{MH} = \sum_k \widehat{v}_k (\widehat{RR}_k) \quad (4.18)$$

with weights

$$\widehat{v}_k = \frac{b_k n_{1k} / N_k}{\sum_\ell b_\ell n_{1\ell} / N_\ell}. \quad (4.19)$$

Because these estimators were derived heuristically, their large sample properties were not derived; nor did Mantel and Haenszel describe an estimate of the large sample variance or the computation of confidence limits.

4.3.2 Test-Based Confidence Limits

One of the first approaches to constructing confidence limits for the assumed common odds ratio, which may also be applied to the Mantel-Haenszel estimate of the common relative risk, is the suggestion by Miettinen (1976) that the aggregate Mantel-Haenszel test be inverted to yield test-based confidence limits. These are a simple generalization of the test-based limits described in Section 2.6.6.

Let $\widehat{\theta}$ denote the log of the Mantel-Haenszel estimate, either $\widehat{\theta} = \log(\widehat{OR}_{MH})$ or $\widehat{\theta} = \log(\widehat{RR}_{MH})$. Then assume that the Mantel-Haenszel test could also be expressed as an estimator-based test. If the variance of the estimate $\sigma_{\widehat{\theta}}^2$ were known, then the test statistic would be constructed using the variance estimated under the null hypothesis of no partial association, that is, using the variance estimate $\widehat{\sigma}_{\widehat{\theta}|H_0}^2$. This would yield a test of the form

$$[Z_{C(MH)}]^2 = X_{C(MH)}^2 = \widehat{\theta}^2 / \widehat{\sigma}_{\widehat{\theta}|H_0}^2. \quad (4.20)$$

Given the observed values $z_{C(MH)}$ and $\hat{\theta}$, the test statistic can be inverted to yield a test-based estimate of the variance of $\hat{\theta}$ as

$$\hat{\sigma}_{\hat{\theta}|H_0}^2 = [\hat{\theta}/z_{C(MH)}]^2. \quad (4.21)$$

This test-based variance could then be used to construct confidence limits for the parameter of interest. For example, for $\hat{\theta} = \log(\widehat{OR}_{MH})$, the resulting test-based $(1 - \alpha)$ -level confidence limits on $\log(OR)$ are provided by

$$\hat{\theta} \pm Z_{1-\alpha/2} \hat{\theta}/z_{C(MH)} \quad (4.22)$$

and those on OR as

$$\begin{aligned} (\widehat{OR}_\ell, \widehat{OR}_u) &= \exp \left[\hat{\theta} \pm Z_{1-\alpha/2} \hat{\theta}/z_{C(MH)} \right] \\ &= \exp \left[\log \left(\widehat{OR}_{MH} \right) \pm \frac{Z_{1-\alpha/2}}{z_{C(MH)}} \log \left(\widehat{OR}_{MH} \right) \right] \\ &= \widehat{OR}_{MH}^{1 \pm \frac{Z_{1-\alpha/2}}{z_{C(MH)}}}. \end{aligned} \quad (4.23)$$

These test-inverted confidence limits are inherently incorrect because they are based on an estimate of the variance derived under the null hypothesis, whereas the proper limits are defined under the alternative hypothesis. However, they often work well in practice. See Halperin (1977) and Greenland (1984), among others, for further discussion of the properties of these confidence limits.

4.3.3 Large Sample Variance of the Log Odds Ratio

In general, it is preferable to obtain asymmetric confidence limits on an odds ratio or relative risk as the exponentiation of the symmetric confidence limits from the log odds ratio or log relative risk. Since the Mantel-Haenszel estimates are weighted averages of the odds ratios and relative risks, not the logs thereof, a large sample variance of the log odds ratio, and also the log relative risk, can be obtained as follows using the δ -method (Hauck, 1979).

Consider the estimate of the common odds ratio OR_{MH} in (4.14), where $\theta = \log(OR_{MH})$. As shown in Problem 2.5.4, given an estimate of the variance of the log odds ratio, $V(\hat{\theta})$, then asymptotically

$$V \left(\widehat{OR}_{MH} \right) \cong (OR_{MH})^2 V(\hat{\theta}) \triangleq \left(\widehat{OR}_{MH} \right)^2 \widehat{V}(\hat{\theta}), \quad (4.24)$$

where " \cong " means "estimated as." However, the Mantel-Haenszel estimate is a weighted average of the stratum-specific odds ratios. If we treat the weights $\{v_k\}$ as known (fixed), then the asymptotic variance $V(\widehat{OR}_{MH})$ can be obtained directly

from (4.15) as

$$\begin{aligned} V(\widehat{OR}_{MH}) &= \sum_k v_k^2 V(\widehat{OR}_k) \cong \sum_k v_k^2 (OR_k)^2 V[\log(\widehat{OR}_k)] \\ &\stackrel{\triangle}{=} \sum_k \widehat{v}_k^2 (\widehat{OR}_k)^2 \widehat{V}[\log(\widehat{OR}_k)]. \end{aligned} \quad (4.25)$$

Substituting into (4.24) yields

$$V(\widehat{\theta}) \cong \frac{V(\widehat{OR}_{MH})}{(OR_{MH})^2} \stackrel{\triangle}{=} \frac{\widehat{V}(\widehat{OR}_{MH})}{(\widehat{OR}_{MH})^2}, \quad (4.26)$$

from which the estimate $\widehat{V}(\widehat{\theta})$ can then be computed. This then yields confidence limits on the log odds ratio and asymmetric confidence limits on the odds ratio. The expressions for the relative risk and log thereof are similar.

Guilbaud (1983) presents a precise derivation of this result, also taking into account the fact that the estimated odds ratio is computed using estimated weights $\{\widehat{v}_k\}$ rather than the true weights. However, from Slutsky's convergence theorem (A.45), since the estimated weights in (4.16) and (4.19) can be shown to converge to constants, the above simple derivation applies with large sample sizes.

Various authors have derived other approximations to the variance of the Mantel-Haenszel estimate of the common log odds ratio $\widehat{\theta} = \log(\widehat{OR}_{MH})$. The most accurate of these is the expression given by Robins, Breslow, and Greenland (1986) and Robins, Greenland, and Breslow (1986). The derivation is omitted. The estimate is based on the following five sums:

$$S_1 = \sum_{k=1}^K a_k d_k / N_k; \quad S_2 = \sum_{k=1}^K b_k c_k / N_k; \quad S_3 = \sum_{k=1}^K (a_k + d_k) a_k d_k / N_k^2; \quad (4.27)$$

$$S_4 = \sum_{k=1}^K (b_k + c_k) b_k c_k / N_k^2; \quad S_5 = \sum_{k=1}^K [(a_k + d_k) b_k c_k + (b_k + c_k) a_k d_k] / N_k^2$$

The estimate of the large sample variance is then calculated in terms of these five sums as follows:

$$\widehat{\sigma}_{\theta}^2 = \widehat{V}[\log(\widehat{OR}_{MH})] = \frac{S_3}{2S_1^2} + \frac{S_5}{2S_1 S_2} + \frac{S_4}{2S_2^2}. \quad (4.28)$$

Example 4.3 Clinical Trial in Duodenal Ulcers (continued)

For this example, Table 4.2 presents a summary of the calculation of the Mantel-Haenszel estimate of the stratified-adjusted odds ratio, and the adjusted relative risk, along with the Hauck estimate of the variance and the corresponding 95% confidence limits. The Mantel-Haenszel estimate of the common odds ratio is $\widehat{OR}_{MH} = 1.634$, and its log is 0.491. The Robins, Breslow and Greenland estimate of the variance

Table 4.2 Relative risks and odds ratios within strata and the Mantel-Haenszel adjusted estimates.

Association Measure	Stratum			Mantel-Haenszel	
	1	2	3	Estimate	95% C.I.
Odds Ratio	0.831	3.750	2.722	1.634	0.90, 2.97
\widehat{v}_k	0.608	0.059	0.333		
$\widehat{V}(\widehat{OR}_{MH})$				0.248	
Relative Risk	0.895	1.688	1.674	1.306	0.95, 1.79
\widehat{v}_k	0.474	0.115	0.411		
$\widehat{V}(\widehat{RR}_{MH})$				0.044	

of $\log(\widehat{OR}_{MH})$ is 0.0813, which yields asymmetric 95% confidence limits for the common odds ratio of (0.934, 2.857). The terms entering into this computation are $S_1 = 15.708$, $S_2 = 9.614$, $S_3 = 9.194$, $S_4 = 4.419$ and $S_5 = 11.709$. For comparison, the test-based confidence limits for the common odds ratio are (0.938, 2.846). Hauck's method yields $\widehat{V}(\widehat{OR}_{MH}) = 0.24792$ with $\widehat{V}(\log \widehat{OR}_{MH}) = 0.09287$, somewhat larger than the Robins, Breslow and Greenland estimate. The resulting confidence limits for OR are (0.899, 2.97), that are somewhat wider.

Compared to the marginal unadjusted analysis, the stratified-adjusted analysis yields a slightly smaller odds ratio in favor of the new drug treatment (1.63 vs. 1.69). Thus, a small part of the original unadjusted estimate of the difference in effectiveness of drug versus placebo was caused by the imbalances in the numbers of subjects from each stratum within each group.

The weights $\{\widehat{v}_k\}$ show that the Mantel-Haenszel estimates give the greatest weight to stratum 1, the least to stratum 2, and more so for the odds ratios than for the relative risks.

4.3.4 Maximum Likelihood Estimate of the Common Odds Ratio

An adjusted estimate of the assumed common odds ratio can also be obtained through maximum likelihood estimation. Under the hypothesis that $E(\widehat{OR}_k) = OR = \varphi$ for all K strata, then the total likelihood is the product of the K stratum-specific conditional hypergeometric likelihoods presented in (2.60)

$$L_c(\varphi) = \prod_{k=1}^K \frac{\binom{n_{1k}}{a_k} \binom{n_{2k}}{m_{1k} - a_k} \varphi^{a_k}}{\sum_{i=a_{\ell k}}^{a_{uk}} \binom{n_{1k}}{i} \binom{n_{2k}}{m_{1k} - i} \varphi^i}, \quad (4.29)$$

where $a_{\ell k}$ and $a_{u k}$ are the limits on the sample space for a_k given the margins in the k th stratum as described in Section 2.4.2. Taking the derivative of $\log L_c(\varphi)$ with respect to φ , the estimating equation for the *MLE* for φ cannot be expressed in a closed-form expression. This approach is described by Birch (1964) and is often called the *conditional maximum likelihood estimate* of the common odds ratio. Example 6.7 describes Newton-Raphson iteration to obtain the *MLE* for φ .

Then in Chapter 7 we also show that the *MLE* of the common log odds ratio can be obtained through a logistic regression model. This latter estimate is often termed the *unconditional MLE* of the common odds ratio.

4.3.5 Minimum Variance Linear Estimators

A third family of estimators is described by Gart (1971) using the principle of weighting inversely proportional to the variances (Meier, 1953), which is derived from weighted least squares estimation. This approach provides an adjusted estimator that is a minimum variance linear estimator (*MVLE*) of θ for measures of association on any "scale" $\theta = G(\pi_1, \pi_2)$ for some smooth function $G(\cdot, \cdot)$. Therefore, these are asymptotically efficient within the class of linear estimators. Since the estimates within each table are consistent, then the *MVLE* is also a consistent estimator and its asymptotic variance is easy to derive.

Using the framework of weighted least squares (Section A.5.3) we have a vector of random variables $\hat{\theta} = (\hat{\theta}_1 \cdots \hat{\theta}_K)^T$, where the assumed model specifies that a common θ applies to all strata such that $E(\hat{\theta}_k) = \theta$ for $k = 1, \dots, K$. Further, the variance of the estimate within the k th stratum is $V(\hat{\theta}_k) = E(\hat{\theta}_k - \theta)^2 = \sigma_{\hat{\theta}_k}^2$, which will be designated simply as σ_k^2 . For each measure of association, these variances are presented in Section 2.3. For now, assume that the $\{\sigma_k^2\}$ are known (fixed). Therefore, under this model, asymptotically

$$\hat{\theta} \cong \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \theta + \varepsilon = \mathbf{J}\theta + \varepsilon, \quad (4.30)$$

where \mathbf{J} is a $K \times 1$ unit vector of ones, and ε is a $K \times 1$ vector where

$$E(\varepsilon) = \mathbf{0}, \quad V(\varepsilon) = \text{diag}(\sigma_1^2 \cdots \sigma_K^2) = \Sigma_{\varepsilon}. \quad (4.31)$$

We express the relationship only asymptotically because $\hat{\theta}_k \xrightarrow{p} \theta_k = \theta$ for all k . Also, $V(\varepsilon_k) = \sigma_k^2$ represents the random sampling variation of the estimate $\hat{\theta}_k$ in the k th stratum about the assumed common parameter θ .

Then from (A.70), the WLS estimate of the common parameter is

$$\hat{\theta} = (\mathbf{J}' \Sigma_{\varepsilon}^{-1} \mathbf{J})^{-1} (\mathbf{J}' \Sigma_{\varepsilon}^{-1} \hat{\theta}). \quad (4.32)$$

Since $\Sigma_{\epsilon}^{-1} = \text{diag}(\sigma_1^{-2} \cdots \sigma_K^{-2})$, this estimator can be expressed as a weighted average of the stratum-specific estimates

$$\hat{\theta} = \frac{\sum_k \sigma_k^{-2} \hat{\theta}_k}{\sum_k \sigma_k^{-2}} = \sum_k \omega_k \hat{\theta}_k, \quad (4.33)$$

where

$$\omega_k = \frac{\sigma_k^{-2}}{\sum_{\ell} \sigma_{\ell}^{-2}} = \frac{\tau_k}{\sum_{\ell} \tau_{\ell}}, \quad (4.34)$$

$\tau_k = \sigma_k^{-2}$, and $\sum_k \omega_k = 1$. Also, from (A.72), the variance of the estimate is

$$V(\hat{\theta}) = \sigma_{\hat{\theta}}^2 = (\mathbf{J}' \Sigma_{\epsilon}^{-1} \mathbf{J})^{-1} = \frac{1}{\sum_k \sigma_k^{-2}}. \quad (4.35)$$

In practice, the estimate is computed using estimated weights $\{\hat{\omega}_k\}$ obtained by substituting the large sample estimate of the stratum-specific variances $\{\hat{\sigma}_k^2\}$ in (4.33) so that

$$\hat{\theta} = \sum_k \hat{\omega}_k \hat{\theta}_k. \quad (4.36)$$

Since $\hat{\sigma}_k^2 \xrightarrow{p} \sigma_k^2$, $\hat{\tau}_k \xrightarrow{p} \tau_k$, and $\hat{\omega}_k \xrightarrow{p} \omega_k$, then from Slutsky's convergence theorem (A.45) the resulting estimate $\hat{\theta}$ is consistent for θ , or $\hat{\theta} \xrightarrow{p} \theta$, and asymptotically $\hat{\theta}$ is distributed as

$$(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma_{\hat{\theta}}^2). \quad (4.37)$$

Likewise, it follows from Slutsky's theorem that a consistent estimate of the variance is obtained by substituting the stratum-specific variance estimates $\{\hat{\sigma}_k^2\}$ into (4.35) so that

$$\hat{V}(\hat{\theta}) = \frac{1}{\sum_k \hat{\sigma}_k^{-2}}. \quad (4.38)$$

This is another derivation of the principle that weighting inversely proportional to the variances provides a minimum variance estimate under an appropriate model. Also, this is a *fixed effects model* because we assume that there is a common value of the parameter for all strata, $E(\hat{\theta}_k) = \theta$ for all k . Therefore, the model assumes that all of the variation *between* the values of the observed $\{\hat{\theta}_k\}$ is caused by random sampling variation about a common value θ .

The explicit expressions for $\hat{\theta}$ and $\hat{V}(\hat{\theta})$ for the risk difference, log odds ratio, and log relative risk are derived in Problem 4.2.3. The following is an example of the required computations.

Example 4.4 Clinical Trial in Duodenal Ulcers (continued)

For the data in Example 4.1, the following is a summary of the computation of the *MVLE* for the stratified-adjusted log odds ratio, where $\hat{\theta}_k = \log(\widehat{OR}_k)$. The

Table 4.3 MVLE of the common parameter θ and its estimated large sample variance for each measure.

Measure	Stratum			Adjusted	
	1	2	3	$\hat{\theta}$	95% C.I.
$\hat{\theta}_k = \text{Risk Difference}$	-0.045	0.306	0.245	0.125	-0.01, 0.26
$\hat{V}(\hat{\theta}_k)$	0.011	0.043	0.010	0.0047	
$\hat{\omega}_k$	0.437	0.110	0.453		
$\hat{\theta}_k = \log \text{Relative Risk}$	-0.111	0.523	0.515	0.281	
$\hat{V}(\hat{\theta}_k)$	0.067	0.167	0.054	0.0254	
$\hat{\omega}_k$	0.376	0.152	0.472		
<i>Relative Risk</i>				1.324	0.97, 1.81
$\hat{\theta}_k = \log \text{Odds Ratio}$	-0.185	1.322	1.001	0.493	
$\hat{V}(\hat{\theta}_k)$	0.188	0.894	0.189	0.0854	
$\hat{\omega}_k$	0.454	0.095	0.451		
<i>Odds Ratio</i>				1.637	0.92, 2.90

estimates, variances, and estimated weights within each stratum are

Stratum	$\hat{\theta}_k$	$\hat{\sigma}_k^2$	$\hat{\sigma}_k^{-2}$	$\hat{\omega}_k$
1	-0.185	0.188	5.319	0.454
2	1.322	0.894	1.118	0.095
3	1.001	0.189	5.277	0.451
<i>Total</i>			11.715	1.0

so that the MVLE is $\hat{\theta} = \log(\widehat{OR}) = (-0.185 \times 0.454) + (1.322 \times 0.095) + (1.001 \times 0.451) = 0.493$ and $\widehat{OR} = 1.637$. The estimated variance of the log odds ratio is $\hat{V}(\hat{\theta}) = 1/11.715 = 0.0854$, which yields an asymmetric 95% C.I. on OR of $\exp(\hat{\theta} \pm 1.96\hat{\sigma}_{\hat{\theta}}) = \exp[0.493 \pm (1.96)\sqrt{0.0854}] = (0.924, 2.903)$. Note that the MVLE estimate and the asymmetric confidence limits are very close to the Mantel-Haenszel estimate and its confidence limits.

Table 4.3 presents a summary of the computations of the MVLE of the common parameter θ and its estimated large sample variance for the risk difference, relative risk, and odds ratio. Note that each of the adjusted estimates is based on a slightly different set of weights. All three estimates, however, give less weight to the second stratum because of its smaller sample size.

Table 4.4 SAS PROC FREQ analysis of Example 4.1.

```

data one;
input k a b c d;
cards;
1 16 20 26 27
2 9 4 3 5
3 28 16 18 28
;
Title1 'Example 4.1: Ulcer Clinical Trial';
data two; set one;
keep i j k f;
*K=Stratum, I=Group, J=Response, F=Frequency;
  i = 1; j = 1; f =a; output;
  i = 2; j = 1; f =b; output;
  i = 1; j = 2; f =c; output;
  i = 2; j = 2; f =d; output;
proc freq; table k*(i j) / chisq nocol nopercent; weight f;
Title2
'Association Of Stratum By Group (k*i) And By Response (k*j)';
proc freq; table k*i*j / cmh; weight f;
Title2 'SAS Mantel-Haenszel Analysis';
run;

```

4.3.6 MVLE Versus Mantel-Haenszel Estimates

The Mantel-Haenszel estimate of the common odds ratio \widehat{OR}_{MH} can be expressed as a linear combination of the stratum-specific odds ratios using weights \widehat{v}_k as shown in (4.15). However, the weights are not proportional to the inverse of the variance of the estimate within each stratum, that is,

$$\widehat{v}_k \not\propto \widehat{\sigma}_k^{-2}, \quad (4.39)$$

where in this context $\sigma_k^2 = V(\widehat{OR}_k)$. Thus, the Mantel-Haenszel estimate is not a minimum variance linear estimator of the common odds ratio and will have a larger variance of the estimate than the *MVLE* of the common odds ratio. Note that in the preceding section we described the *MVLE* of the common log odds ratio, which is preferable for the purpose of constructing confidence intervals. However, the *MVLE* of the common odds ratio, without the log transformation, may also be readily obtained.

Nevertheless, the Mantel-Haenszel estimate still has a favorable total mean square error (*MSE*) compared to the *MVLE* and the *MLE* because the individual odds ratios are not unbiased with finite samples, and thus the adjusted estimators are also not unbiased. From the principle of partitioning of variation in Section A.1.3, (A.5), the total *MSE* of an estimator can be partitioned as $MSE(\widehat{\theta}) = V(\widehat{\theta}) + Bias^2$. Gart

Table 4.5 SAS PROC FREQ Mantel-Haenszel analysis of Example 4.1.
 SUMMARY STATISTICS FOR I BY J
 CONTROLLING FOR K

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	3.005	0.083
2	Row Mean Scores Differ	1	3.005	0.083
3	General Association	1	3.005	0.083

Estimates of the Common Relative Risk (Row1/Row2)

95%

Type of Study	Method	Value	Confidence	Bounds
Case-Control (Odds Ratio)	Mantel-Haenszel	1.634	0.938	2.846
	Logit	1.637	0.924	2.903
Cohort (Col1 Risk)	Mantel-Haenszel	1.306	0.966	1.767
	Logit	1.324	0.969	1.810
Cohort (Col2 Risk)	Mantel-Haenszel	0.796	0.615	1.030
	Logit	0.835	0.645	1.083

The confidence bounds for the M-H estimates are test-based.

Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square = 4.626 DF = 2 Prob = 0.099

(1971), McKinlay (1978), Breslow (1981), and Hauck (1989), among others, have shown that the *MSE* of the Mantel-Haenszel estimate is close to that of the *MLE* in various settings. Thus, there is a trade-off between the slightly larger variance of the Mantel-Haenszel estimate and its slightly smaller bias with finite samples than is the case with the *MVLE* estimates.

4.3.7 SAS PROC FREQ

Some, but not all, of the preceding methods are provided by the SAS procedure PROC FREQ for the analysis of cross-classified frequency data. Section 2.7 describes the SAS output and computations for a single 2×2 table. This procedure also conducts

a Cochran-Mantel-Haenszel analysis of multiple 2×2 tables. For illustration, Table 4.4 presents SAS statements to conduct an analysis of the data from Example 4.1. The results are presented in Table 4.5. Note that a data set is required that includes the level of stratum (k), group (i), and response (j) along with the frequency (f) within each cell of each 2×2 table.

The first call of PROC FREQ generates the tables that assess the association of the strata with group membership and with the response. These analyses are discussed in the following section. Thus, these pages of SAS output are not displayed. The second call of PROC FREQ conducts the analysis of the association between group and response within each stratum, and the Mantel-Haenszel analysis over strata. The 2×2 tables of group by response within each stratum also are not displayed because this information is presented in Section 4.2. The Mantel-Haenszel stratified-adjusted analysis is shown in Table 4.5.

The Mantel-Haenszel analysis first presents three stratified-adjusted tests; the test of nonzero correlation, the test that the row mean scores differ, and the test of general association. For multiple 2×2 tables, all three tests are equivalent to the Mantel-Haenszel test described herein. For $R \times C$ tables with $R > 2$ or $C > 2$, the three tests differ (see Stokes et al., 2000).

This analysis is followed by measures of association. SAS labels these as measures of relative risk. Three measures are computed: *case-control*, *cohort (column 1 risk)* and *cohort (column 2 risk)*. As described in Section 2.7, these refer to the odds ratio, column 1 relative risk, and the column 2 relative risk, respectively. For each, the Mantel-Haenszel estimate and a "logit" estimate are computed. For the odds ratio (case-control relative risk) the logit estimate refers to the *MVLE*, actually the exponentiation of the *MVLE* of the common log odds ratio. The logit confidence limits are the asymmetric confidence limits for the odds ratio obtained from the *MVLE* confidence limits for the common log odds ratio. For the column 1 and the column 2 relative risks, the Mantel-Haenszel estimate of the stratified-adjusted relative risk and the test-inverted confidence limits are presented. The *MVLE* of the common relative risk and the corresponding asymmetric confidence limits are also presented. These are also labeled as the *logit* estimates, although they are based on the log transformation, not the logit.

4.4 NATURE OF COVARIATE ADJUSTMENT

For whichever measure is chosen to describe the association between the treatment or exposure group and the response, the marginal unadjusted estimate of the parameter $\hat{\theta}_0$ is often different from the stratified-adjusted estimate $\hat{\theta}$. Also, the value of the marginal unadjusted chi-square test, and its level of significance, are often different from those in the stratified-adjusted analysis. Part of this phenomenon may be explained by simple sampling variation, but a more important consideration is that the parameters in the conditional within-stratum analysis (the $\{\theta_k\}$), the marginal unadjusted analysis (θ_0), and the stratified-adjusted analysis (θ) are all distinct. Thus, in general, one expects that the marginal parameter in the population will differ

conceptually, and thus in value, from the assumed common value among a defined set of strata, that is, $\theta_* \neq \theta$.

The marginal parameter θ_* , such as the odds ratio, is the expected value of the sample estimate on sampling n_1 and n_2 observations from a large (infinite) population. Within the k th of K strata, for K fixed, the stratum-specific value θ_k is the expectation $E(\hat{\theta}_k)$ on sampling n_{1k} and n_{2k} observations from the large population defined by stratum k . In the stratified-adjusted model, it is assumed that the stratum-specific parameters share a common value θ that is distinct from θ_* . In general, therefore, one should expect some difference between the marginal unadjusted analysis and the stratified-adjusted analysis.

Thus, there is no unique adjusted estimate of the parameter of interest nor a unique p -value for a statistical test that may be considered "correct," or "the truth." The results obtained depend on the model employed, which in this case refers to the scale used to measure the effect (*RD*, *RR*, *OR*, etc.), and on the set of covariates used to define the strata. Therefore, the conclusion of the analysis may depend on the measure chosen to reflect the difference between groups and the covariate(s) chosen to adjust for. In fact, this is true of all models.

4.4.1 Confounding and Effect Modification

When epidemiologists contrast the marginal unadjusted estimate versus the stratum-specific estimates versus the stratified-adjusted estimate, usually in terms of the odds ratio or relative risk, they often draw a distinction between the influence of a confounding variable versus effect modification. See Schlesselman (1982), Kleinbaum et al. (1982), Rothman (1986) and Kelsey et al. (1996), among many, for further discussion.

Confounding usually refers to the situation where the stratification variable is the true causal agent and the treatment or exposure group (the independent variable) is indirectly related to the outcome through its association with the causal stratification variable. For example, individuals who drink large amounts of coffee have a much higher prevalence of smoking than is found in the general population. Thus, any association between the amount of coffee drunk per day and the risk of heart disease or cancer will likely be confounded with the association between smoking and coffee drinking. In this case, one expects that the unadjusted association between coffee drinking and heart disease is different from the adjusted association after stratification by smoking versus not smoking. In fact, one expects the adjusted odds ratio to be substantially smaller. Therefore, it is now common to refer to *confounding* whenever an adjustment for another variable (the covariate) leads to a change in the nature of the association between the independent variable and the response. Such informal usage, however, is not precise unless the covariate can in fact be viewed as a true causal agent.

In some cases the stratification adjustment may result in an increase in the strength of association between the independent and dependent variables. In extreme cases this is referred to as *Simpson's paradox*. However, this is a misnomer. In multiple re-

gression, for example, it is well known that adjustment for other covariates may result in a substantial increase in the strength of the association between the independent variable and the response.

Confounding is an example of an *antagonistic* effect of adjustment where some of the association of the independent variable with the response is explained by the association between the independent variable and the covariate and between the covariate and the response. However, it is also possible that the adjustment may introduce a *synergistic* effect between the independent variable and the covariate such that the covariate-adjusted association between the independent variable and the response is greater than the marginal unadjusted association.

Basically, this can be viewed as follows: Consider the multiple regression case where all variates are quantitative. Let X be the independent variable, Y the dependent variable (response), and Z the covariate. If Z has a strong association with X but a weak association with Y , then including Z in the model will help explain some of the residual error or noise in the measurements of X , thus allowing the signal in X to better "predict" or be more strongly correlated with the values of Y . In the social sciences, the covariate Z in this case would be called a *suppressor variable*.

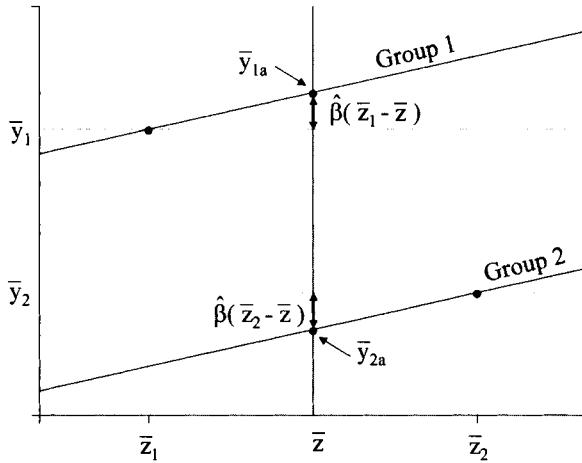
Effect modification, on the other hand, refers the observation that the stratum-specific estimates, the $\{\theta_k\}$, differ among the specific strata. Subsequently this is referred to as *heterogeneity* among strata, or an *interaction* between the group and strata effects. For example, assume that the association between body weight and the risk of developing diabetes were found to differ significantly between men and women, such as odds ratios of 1.7 per 5 kg greater weight for women and 2.8 for men, where the difference between men and women (between strata) is statistically significant. In this case we would say that gender modifies or interacts with the effect of body weight on the risk of developing diabetes.

4.4.2 Stratification Adjustment and Regression Adjustment

Insight into the nature of covariate adjustment is also provided by the analysis of covariance (ANCOVA). Consider that we have two groups represented by the independent variable $X = 1$ or 2 and we wish to estimate the difference between groups in the means of a quantitative dependent variable Y . However, we also wish to adjust this estimate for any differences between groups with respect to a quantitative covariate Z based on the regression of Y on Z . The elements of an analysis of covariance are depicted in Figure 4.1. Assume that the slope of the relationship of the response (Y) with the covariate (Z) is the same in both groups so that the regression lines are parallel. By chance, or because of confounding, the mean value of the covariate may differ between the two groups, such as $\bar{z}_1 < \bar{z}_2$. Thus, some of the difference between the unadjusted Y means in the two groups ($\hat{\theta}_u = \bar{y}_1 - \bar{y}_2$) is attributable to a bias introduced by the difference in the covariate means ($\bar{z}_1 - \bar{z}_2$) between the groups.

Based on the estimated slope of the regression of Y on Z , that is assumed to be the same in the two groups, the bias introduced into the estimate of each Y mean can

Fig. 4.1 Covariance adjustment of the mean value in each group (\bar{y}_i) by removing the bias $\hat{\beta}(\bar{z}_i - \bar{z})$ caused by the imbalance in the covariate means (\bar{z}_i), $i = 1, 2$.



be estimated and then removed. Usually, this is expressed by assessing the expected difference between groups had both groups had the same mean for Z , that is, when $\bar{z}_1 = \bar{z}_2 = \bar{z}$. This adjusted difference is provided by $\hat{\theta}_a = \bar{y}_{1a} - \bar{y}_{2a}$, where the adjusted Y means \bar{y}_{1a} and \bar{y}_{2a} are the values along the regression lines for which $z = \bar{z}$. Thus, in the i th group,

$$\bar{y}_{ia} = \bar{y}_i - (\bar{z}_i - \bar{z})\hat{\beta}, \quad (4.40)$$

and the difference between each unadjusted mean (\bar{y}_i) and the adjusted mean (\bar{y}_{ia}) is the magnitude of the estimated bias introduced by the difference in the covariate mean from its overall mean ($\bar{z}_i - \bar{z}$) based on the estimated slope $\hat{\beta}$ that is assumed to be the same in both groups.

The assumption of a common slope between groups implies that the difference between the conditional expectations $\theta_z = E(y_1|z) - E(y_2|z)$ is constant for all values of the covariate Z . This is also called the assumption of parallelism, or of homogeneity of the group difference for all values of the covariate. This is directly equivalent to the assumption of a common difference between groups on some scale when z takes on discrete values rather than being continuous. Thus, for a discrete covariate, an adjustment using a regression model is conceptually equivalent to an adjustment using direct stratification, although the methods of estimation may differ, for example, iterative maximum likelihood for a logistic regression model, and one-step weighted least squares for the *MVLE* estimates of a common odds ratio. Later in Chapter 7, we show more directly the asymptotic equivalence between a stratification-

adjusted Mantel-Haenszel analysis and an adjustment for the same covariate using a logistic regression model.

It should be noted that the above covariate adjustment may also be obtained from an analysis of the residuals of the relationship of Y on Z . Assuming a quantitative covariate Z with estimated common slope $\hat{\beta}$ in each group, then the residual for the j th observation in the i th group is

$$e_{ij} = y_{ij} - z_{ij}\hat{\beta}. \quad (4.41)$$

Thus, the adjusted difference between groups is also provided by the difference in the group means of the residuals $\hat{\theta}_a = \bar{y}_{1a} - \bar{y}_{2a} = \bar{e}_1 - \bar{e}_2$, the constant $\bar{z}\hat{\beta}$ canceling.

4.4.3 When Does Adjustment Matter?

Further insight into the nature of covariate adjustment is provided by contrasting a simple versus a partial correlation coefficient for quantitative observations, as in multiple regression. Again consider three quantitative variates X , Y , and Z , analogous to group, response, and the stratifying covariate, respectively. The unadjusted correlation ρ_{xy} is a measure of the marginal association between the independent (X) and dependent (Y) variables, without consideration of the association between either variable with the possible mediating covariate (Z). The partial or adjusted correlation, however, examines the association between X and Y after removing the association of each with Z . The partial correlation can be expressed as

$$\rho_{xy,z} = \text{corr} [e(x|z), e(y|z)], \quad (4.42)$$

where $e(x|z) = x - E(x|z)$ is the residual of x from its conditional expectation given the value z for each observation, and $e(y|z) = y - E(y|z)$ is the corresponding residual of y given the value z . Algebraically, the partial correlation reduces to

$$\rho_{xy,z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1 - \rho_{xz}^2)(1 - \rho_{yz}^2)}}, \quad (4.43)$$

where ρ_{yz} is the correlation between the mediating or stratification variable Z and the response Y , and where ρ_{xz} the correlation between the independent variable X and Z . Thus, $\rho_{xy,z} = \rho_{xy}$ if and only if both $\rho_{xz} = 0$ and $\rho_{yz} = 0$, that is, if there is no association between the covariate and the independent variable and between the covariate and the response. If either $\rho_{xz} \neq 0$ or $\rho_{yz} \neq 0$, then $\rho_{xy,z} \neq \rho_{xy}$.

Cochran (1983) illustrated an equivalent relationship in a stratified analysis of 2×2 tables with only two strata. As in Section 3.6.4, assume that there is a constant risk difference θ within each stratum, but the probability of the positive outcome in the control group differs between strata, $\pi_{21} \neq \pi_{22}$. Thus, the exposed or treated group probability in the k th stratum is $\pi_{1k} = \theta + \pi_{2k}$ for $k = 1, 2$. Also assume that the expected sample fraction in the k th stratum is $\zeta_k = E(N_k/N)$ for $k = 1, 2$. We also generalize the model beyond that in Section 3.6.4 to allow unequal treatment group sample fractions among strata such that within the k th stratum the expected

sample fractions within each group are $\xi_{ik} = E(n_{ik}/N_k)$ ($i = 1, 2$; $k = 1, 2$). It is instructive to summarize the model as follows:

Stratum	Stratum Fraction	Group Sample Fractions		Probability +	
		Exposed	Control	Exposed	Control
1	ζ_1	ξ_{11}	ξ_{21}	$\theta + \pi_{21}$	π_{21}
2	ζ_2	ξ_{12}	ξ_{22}	$\theta + \pi_{22}$	π_{22}

Now let κ_i refer to the fraction of all subjects in the i th group who are from the first stratum ($k = 1$), where

$$\kappa_i = \frac{\xi_{i1}\zeta_1}{\xi_{i1}\zeta_1 + \xi_{i2}\zeta_2} \quad (4.44)$$

for $i = 1, 2$. Then the expected difference in the marginal unadjusted analysis is

$$\begin{aligned} E(p_{1\bullet} - p_{2\bullet}) &= \frac{(\theta + \pi_{21})\xi_{11}\zeta_1 + (\theta + \pi_{22})\xi_{12}\zeta_2}{\xi_{11}\zeta_1 + \xi_{12}\zeta_2} - \frac{\pi_{21}\xi_{21}\zeta_1 + \pi_{22}\xi_{22}\zeta_2}{\xi_{21}\zeta_1 + \xi_{22}\zeta_2} \\ &= (\theta + \pi_{21})\kappa_1 + (\theta + \pi_{22})(1 - \kappa_1) - \pi_{21}\kappa_2 - \pi_{22}(1 - \kappa_2) \\ &= \theta + (\kappa_1 - \kappa_2)(\pi_{21} - \pi_{22}). \end{aligned} \quad (4.45)$$

Therefore, the marginal unadjusted estimate of the risk difference is biased by the quantity $(\kappa_1 - \kappa_2)(\pi_{21} - \pi_{22})$, where $\kappa_1 - \kappa_2$ reflects the degree of exposure or treatment group imbalance between the strata, and where $\pi_{21} - \pi_{22}$ reflects the differential in risk between strata. Conversely, a stratified analysis using a weighted average of the differences between groups is unbiased since within each stratum, $E(p_{1k} - p_{2k}) = \theta$ for $k = 1, 2$.

For example, consider a study with the following design elements:

Stratum	Stratum Fraction	Group Sample Fractions		Probability +	
		Exposed	Control	Exposed	Control
1	0.30	0.75	0.25	0.35	0.20
2	0.70	0.40	0.60	0.55	0.40

where $\kappa_1 = (0.3 \times 0.75)/[(0.3 \times 0.75) + (0.7 \times 0.4)] = 0.4455$ and $\kappa_2 = 0.1515$. Because of the imbalance in the treatment group sample fractions and the difference in risk between strata, the difference in the marginal proportions from the unadjusted analysis will be biased. In this case, the bias of the marginal test statistic is $(0.4455 - 0.1515) \times (0.2 - 0.4) = -0.059$, which is substantial relative to the true difference of 0.15 within each stratum.

Beach and Meier (1989), Canner, (1991) and Lachin and Bautista (1995), among others, demonstrated that the same relationships apply to other marginal unadjusted measures of association, such as the log odds ratio. Therefore, the difference between the unadjusted measure of association between group and response, equivalent to ρ_{xy} , and the stratified-adjusted measure of association, equivalent to $\rho_{xy,z}$, is a function of the degree of association between the covariate with group membership and with the outcome. In a stratified analysis, these associations are reflected by the $K \times 2$ contingency table for strata by group and the $K \times 2$ table for stratum by response.

Example 4.5 *Clinical Trial in Duodenal Ulcers (continued)*

The following tables describe the association between the stratification covariate ulcer type and treatment group (drug vs. placebo) and the association between the covariate and the likelihood of healing (+) versus not (-):

Stratum	Stratum by Group				Stratum by Response			
	#		%		#		%	
	D	P	D	P	+	-	+	-
1	42	47	47.2	52.8	36	53	40.5	59.5
2	12	9	57.1	42.9	13	8	61.9	38.1
3	46	44	51.1	48.9	44	46	48.9	51.1

These tables would be provided by the SAS statements in Table 4.4.

In the table of stratum by group, if there are no group imbalances, then the same proportion of subjects should be from each group within each stratum. However, proportionately fewer patients from the drug-treated group fall in the first stratum (47.2%) and more in the second stratum (57%). The contingency test of association for this table is $X^2 = 0.754$ on 2 df , $p \leq 0.686$. Similarly, in the table of stratum by response, if the covariate was not associated with the likelihood of healing, that is, was not a risk factor for healing, the same proportion of patients would have a "+" response within each stratum. However, there is proportionately less healing in the first stratum (40.5%) and more in the second stratum (61.9%). For this table, $X^2 = 3.519$ on 2 df , $p \leq 0.172$.

Neither table shows a significant association with the stratification covariate, although in this regard statistical significance is less important than the degree of proportionate differences because significance will also be a function of the sample size. These proportionate differences are relatively minor, so the covariate-adjusted analysis differs only slightly from the unadjusted analysis. This should be expected in a randomized study because, through randomization, it is unlikely that a substantial imbalance in the treatment group proportions among strata will occur. The following are additional examples of nonrandomized studies in which the adjustment does make a difference.

Example 4.6 *Religion and Mortality*

Zuckerman et al. (1984) present the following stratified analysis of a prospective cohort study of the association between having religious beliefs versus not on mortality over a period of two years among an elderly population that was forced to relocate to nursing homes. The four strata comprised (1) healthy males, (2) ill males, (3) healthy females, and (4) ill females. Note that the strata are simultaneously adjusting for gender and for healthy versus ill. The 2×2 tables for the four strata, in abbreviated

Table 4.6 Stratified-adjusted analysis of the data from Example 4.6.

Measure	Stratum				Marginal	Adjusted	
	$\hat{\theta}_k$	1	2	3	4	$\hat{\theta}_*$	$\hat{\theta}$
RD		-0.0048	-0.2288	-0.0098	-0.0904	-0.0785	-0.0265
$\hat{V}(\hat{\theta}_k)$		0.0054	0.0152	0.0008	0.0049	0.0011	0.0006
$\log RR$		-0.0408	-0.7892	-0.3615	-0.6016	-0.6685	-0.5303
$\hat{V}(\hat{\theta}_k)$		0.3976	0.2471	0.9726	0.2002	0.0799	0.0795
RR		0.9600	0.4542	0.6966	0.5480	0.5125	0.5885
$\log OR$		-0.0462	-1.1215	-0.3716	-0.7087	-0.7580	-0.6297
$\hat{V}(\hat{\theta}_k)$		0.5093	0.4413	1.0282	0.2793	0.1017	0.1139
OR		0.9548	0.3258	0.6897	0.4923	0.4686	0.5327

notation, are

Stratum	Religious			Nonreligious		
	a_k	n_{1k}	p_{1k}	b_k	n_{2k}	p_{2k}
1: Healthy females	4	35	0.114	5	42	0.119
2: Ill females	4	21	0.190	13	31	0.419
3: Healthy males	2	89	0.022	2	62	0.032
4: Ill males	8	73	0.110	9	45	0.200
<i>Marginal</i>	18	218	0.083	29	180	0.161

where a_k and b_k , respectively, are the numbers of patients who died in the religious and nonreligious groups (reproduced with permission). The measures of association within each stratum, for the pooled data marginally with no stratification adjustment, and the *MVLE* stratified-adjusted estimates and their variances are presented in Table 4.6. For all measures of association, the magnitude of the group effect in the marginal unadjusted analysis is greater than that in the stratified-adjusted analysis; that is, the $\hat{\theta}$ is closer to zero than is the $\hat{\theta}_*$.

Likewise, the Mantel-Haenszel analysis of the odds ratios, compared to the marginal unadjusted analysis, is presented in Table 4.7. The stratified-adjusted Mantel-Haenszel estimate is close to that provided by the *MVLE* and also shows less association between religious versus nonreligious with mortality. The net result is that whereas the unadjusted Mantel-Haenszel test is significant at $p \leq 0.02$, the adjusted statistic is not significant at the usual 0.05 significance level.

To determine why the stratified-adjusted analysis provides a somewhat different conclusion from the unadjusted analysis, it is necessary to examine the degree of association of the stratification variables with group membership and with mortality.

Table 4.7 Mantel-Haenszel analysis of odds ratios for the data in Example 4.6.

Mantel-Haenszel Analysis		
	Marginal Unadjusted	Stratified- Adjusted
\widehat{OR}	0.4686	0.5287
$\log \widehat{OR}$	-0.758	-0.637
$\widehat{V} [\log \widehat{OR}]$	0.1017	0.1116
95% C.I. for OR	0.251, 0.875	0.275, 1.018
X^2_{MH}	5.825	3.689
$p \leq$	0.0158	0.0548

The corresponding 4×2 tables are

Stratum	By Group			By Response		
	R	\bar{R}	%R	#	Died	Alive
1: Healthy females	35	42	45.5	9	68	11.7
2: Ill females	21	31	40.4	17	35	32.7
3: Healthy males	89	62	58.9	4	147	2.7
4: Ill males	73	45	61.9	17	101	14.4

The chi-square test of association between stratum and group membership is $X^2 = 10.499$ on 3 df with $p \leq 0.015$. The proportion of religious subjects within strata 3 and 4 (all males) is higher than that in strata 1 and 2 (all females), so that the effect of religious versus nonreligious is somewhat associated with the effect of gender. The test of association between stratum and mortality is also highly significant ($X^2 = 34.706$, $p \leq 0.001$). The mortality among ill patients of both genders (strata 2 and 4) is much higher than that among healthy patients (strata 1 and 3), and the mortality among females is greater than that among males.

Thus, some of the association between religious versus nonreligious in the marginal analysis is explained by the imbalances in the proportions of religious subjects among the four strata and the corresponding differences in mortality between strata. The stratified-adjusted analysis eliminates this "confounding" by comparing the religious versus nonreligious subjects within strata and then averaging these differences over strata. Mantel and Haenszel (1959) refer to this as the principle of comparing like-to-like in the stratified analysis.

Example 4.7 Simpson's Paradox

In some cases the stratification adjustment yields not only an adjustment in the magnitude of the association between the independent variable and the response (the quantity of the effect), but also in the quality or direction of the association (the

quality of the effect). *Simpson's paradox* refers to cases where marginally there is no group effect but after adjustment there is a big group effect. This is illustrated by the following hypothetical data from Stokes et al. (2000) of a health policy opinion survey of the association between stress (no/yes) and the risk (probability) of favoring a proposed new health policy among residents of urban and rural communities.

The 2×2 tables for the two strata and marginally are

Stratum	Not-Stressed			Stressed		
	a_k	n_{1k}	p_{1k}	b_k	n_{2k}	p_{2k}
1: Urban	48	60	0.800	96	190	0.505
2: Rural	55	190	0.289	7	60	0.117
Marginal	103	250	0.412	103	250	0.412

where a_k and b_k are the numbers favoring the new health policy in each stress group (reproduced with permission). The marginal unadjusted analysis shows equal proportions favoring the new policy so that the unadjusted odds ratio is 1.0. However, when stratified by urban versus rural residence the following odds ratios and *MVLE* of the common odds ratio $\hat{\theta}$ are obtained.

Measure	Stratum		
	1	2	$\hat{\theta}$
$\hat{\theta}_k = \log Odds ratio$	1.365	1.126	1.270
$\hat{V}(\hat{\theta}_k)$	0.125	0.187	0.075
Odds ratio	3.917	3.085	3.559

Note that the n_{1k} and n_{2k} are reversed in the two strata so that there is a large group imbalance between strata. Among the urban subjects, 76% are stressed versus only 24% among the rural subjects, $X^2 = 135.2$ on 1 df , $p \leq 0.001$. Also, 57.6% of the urban subjects favored the new health policy versus 24.8% among rural subjects, $X^2 = 55.5$ on 1 df , $p \leq 0.001$. Comparing like-to-like within strata, the resulting odds ratios within the two strata are 3.917 and 3.085, so that the stratified-adjusted *MVLE* of the odds ratio is 3.559. The Mantel-Haenszel test statistic is also highly significant, $X^2_{MH} = 23.05$, $p \leq 0.0001$.

Thus, we observe a strong positive association within strata, which yields a significant association when adjusted over strata, whereas marginally absolutely no association is seen. This is an illustration of a synergistic effect where the covariate adjustment enhances estimation of the association between the independent and dependent variables.

4.5 MULTIVARIATE TESTS OF HYPOTHESES

4.5.1 Multivariate Null Hypothesis

The Mantel-Haenszel test is, in fact, a multivariate test of the joint multivariate null hypothesis of no association in any of the K strata versus a *restricted* alternative

hypothesis that a nonzero common log odds ratio applies to all strata. However, there are other multivariate tests that could be constructed that would provide greater power against other alternative hypotheses.

In a stratified analysis with K strata, we wish to conduct a test for a vector of K association parameters $\boldsymbol{\theta} = (\theta_1 \cdots \theta_K)^T$. The $\{\theta_k\}$ can be measured on any scale of our choosing such as $\theta_k = G(\pi_{1k}, \pi_{2k})$ for some differentiable function $G(\cdot, \cdot)$. The vector of sample estimates is assumed to be asymptotically distributed as

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}_K(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}) \quad (4.46)$$

or

$$\begin{bmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2 \\ \vdots \\ \hat{\theta}_K - \theta_K \end{bmatrix} \xrightarrow{d} \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}), \quad (4.47)$$

with a known (or consistently estimable) covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = V(\hat{\boldsymbol{\theta}})$. For the case of K 2×2 tables, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = \text{diag}(\sigma_{\hat{\theta}_1}^2 \cdots \sigma_{\hat{\theta}_K}^2)$, which is by definition positive definite. Under the null hypothesis $H_0: \pi_{1k} = \pi_{2k} = \pi_k$, let θ_0 designate the null value $\theta_k = \theta_0 = G(\pi_k, \pi_k)$ for all k . The equivalent null hypothesis in terms of the values of $\{\theta_k\}$ is

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_K = \theta_0 \quad \text{or} \quad H_0: \boldsymbol{\theta} = \mathbf{J}\theta_0, \quad (4.48)$$

which is a joint multivariate null hypothesis for the vector $\boldsymbol{\theta}$, where \mathbf{J} is the unit vector, as in (4.30). A variety of statistical tests can be employed to test H_0 , each of which will be optimal against a specific type of alternative hypothesis.

4.5.2 Omnibus Test

The omnibus alternative hypothesis specifies that at least one of the elements of $\boldsymbol{\theta}$ differs from the null value, or

$$H_{1O}: \theta_k \neq \theta_0 \text{ for some } k, 1 \leq k \leq K, \quad \text{or} \quad H_{1O}: \boldsymbol{\theta} \neq \mathbf{J}\theta_0. \quad (4.49)$$

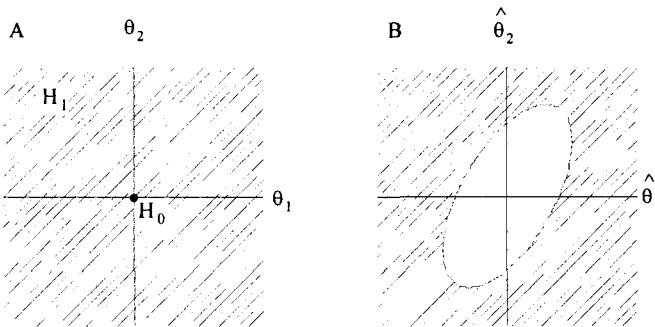
Thus, H_0 specifies that there is no association between group and response in any table, versus H_1 that there is some association in either direction in at least one table, that is, perhaps favoring either group 1 or group 2 for any table.

The asymptotically most powerful test of H_0 versus H_{1O} is the T^2 -like test statistic attributed to Wald (1943)

$$X_O^2 = (\hat{\boldsymbol{\theta}} - \mathbf{J}\theta_0)' \boldsymbol{\Sigma}_0^{-1} (\hat{\boldsymbol{\theta}} - \mathbf{J}\theta_0), \quad (4.50)$$

which uses the covariance matrix defined under the null hypothesis, $\boldsymbol{\Sigma}(\theta_0) = \boldsymbol{\Sigma}_0$, and is asymptotically distributed as χ_K^2 under H_0 . In practice, a consistent estimate of the covariance matrix $\hat{\boldsymbol{\Sigma}}_0$ is employed.

Fig. 4.2 The null (H_0) and alternative (H_1) hypotheses for the omnibus test with two parameters $\theta = (\theta_1, \theta_2)$ (A) and the corresponding rejection region (B).



To illustrate the nature of this test, consider the case of a bivariate statistic ($K = 2$), or the case of two strata, where we wish to test $H_0: \theta_1 = \theta_2 = \theta_0 = 0$. Figure 4.2A describes the null and alternative hypotheses (H_0 and H_{1O}) for this test. The alternative parameter space is *omnibus* or all-inclusive and consists of all points in the two-dimensional space other than the origin.

To describe the rejection region for the test, consider the case of a bivariate statistic vector with correlated elements where the elements of Σ_0 are

$$\Sigma_0 = \frac{1}{N} \begin{bmatrix} 1 & 0.9428 \\ 0.9428 & 2 \end{bmatrix}$$

such that the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ is $2/3$, and the variance of $\hat{\theta}_1$ is half that of $\hat{\theta}_2$. In a stratified analysis, the estimates are independent, and thus uncorrelated. However, it is instructive to consider the more general case of correlated estimates as may apply to the analysis of repeated measures, as one example.

The rejection region for the omnibus test in this instance is shown in Figure 4.2B for a sample of $N = 100$ and a Type I error probability of $\alpha = 0.05$. The rejection region for this test is defined by an ellipse that specifies all values of $\hat{\theta}_1$ and $\hat{\theta}_2$ such that the test statistic $X_O^2 = \hat{\theta}' \hat{\Sigma}_0^{-1} \hat{\theta} = \chi^2_{2(0.95)} = 5.991$. The rejection ellipse is defined along the 45° axis of values for $\hat{\theta}_1$ and $\hat{\theta}_2$ such that the longevity of the ellipse is determined by the relative variances of $\hat{\theta}_1$ and $\hat{\theta}_2$ and the direction of their correlation (positive in this case). All points $\hat{\theta}_1$ and $\hat{\theta}_2$ interior to the ellipse lead

to failure to reject the null hypothesis; those on or exterior to the ellipse lead to rejection of H_0 in favor of H_{1O} that a difference from zero of some magnitude in either direction exists for at least one of the values θ_1 or θ_2 . Thus, the statistic can lead to rejection of H_0 when $\hat{\theta}_1$ is positive, say favoring group 2, and $\hat{\theta}_2$ is negative, favoring group 1.

In the case of K stratified 2×2 tables, since Σ_0 is a diagonal matrix, then the computation of the omnibus test X_O^2 in (4.50) simplifies to

$$X_O^2 = \sum_k \frac{(\hat{\theta}_k - \theta_0)^2}{\sigma_{0k}^2}, \quad (4.51)$$

where $\sigma_{0k}^2 = V(\hat{\theta}_k | H_0)$. In this setting, $\theta_0 = 0$ for θ defined as the risk difference, the log relative risk, or the log odds ratio.

When expressed in terms of the risk difference $\theta_k = \pi_{1k} - \pi_{2k}$, then under H_0 : $\theta_0 = 0$,

$$V_{0k} = \sigma_{0k}^2 = \pi_k (1 - \pi_k) \left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right). \quad (4.52)$$

Based on the consistent estimate $\hat{\theta}_k = p_{1k} - p_{2k}$, the variance is consistently estimated as

$$\hat{V}_{0k} = \hat{\sigma}_{0k}^2 = p_k (1 - p_k) \left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right), \quad (4.53)$$

with $p_k = m_{1k}/N_k$ (see Section 2.3.1). Using the simpler notation \hat{V}_{0k} to refer to $\hat{\sigma}_{0k}^2$,

$$X_O^2 = \sum_k \frac{(p_{1k} - p_{2k})^2}{\hat{V}_{0k}} = \sum_k X_k^2, \quad (4.54)$$

where X_k^2 is the Pearson contingency (Cochran) X^2 value for the k th stratum, which is asymptotically distributed as χ_K^2 on K df.

In Problem 2.9 we showed that the Z -test (and thus the chi-square test) for a single 2×2 table is asymptotically invariant to the choice of scale. Thus, the omnibus test using risk differences, log relative risks, or log odds ratios are all asymptotically equivalent. In practice, therefore, the test is usually computed only in terms of the risk differences.

4.5.3 Multiple Tests

Another approach to a simultaneous test of the K -parameter joint null hypothesis H_0 is to conduct K separate tests and to employ a correction for multiple tests as described in Section 2.11.4. However, for statistics that are jointly distributed as multivariate normal, such as here, multiple test procedures are still conservative compared to the T^2 -type multivariate test.

Example 4.8 *Religion and Mortality (continued)*

For the four strata in Example 4.6, the multivariate omnibus test in terms of the risk differences $\theta_k = \pi_{1k} - \pi_{2k}$ provides a test of $H_0: [\theta_1 = 0, \theta_2 = 0, \theta_3 = 0, \theta_4 = 0]$ against the alternative $H_{1O}: \theta_k \neq 0$ for some $1 \leq k \leq 4$. The within-stratum values of the contingency (Cochran) 1 df X^2 tests are 0.00419, 2.98042, 0.13571, and 1.84539. Thus, the omnibus test is $X_O^2 = \sum_k X_k^2 = 4.9657$ on 4 df , with $p \leq 0.291$ that is not statistically significant. This is in contrast to the significant unadjusted analysis and the nearly significant stratified-adjusted Mantel-Haenszel analysis.

Example 4.9 *Clinical Trial in Duodenal Ulcers (continued)*

Similarly, for Example 4.1, the omnibus test is $X_O^2 = 0.18299 + 2.03606 + 5.40486 = 7.6239$ on 3 df with $p \leq 0.0545$ that approaches significance at the 0.05 level. The p -value for this omnibus test is smaller than that for the unadjusted and Mantel-Haenszel stratified-adjusted test of a common odds ratio.

4.5.4 Partitioning of the Omnibus Alternative Hypothesis

In general, the omnibus test and the Mantel-Haenszel test differ because of the difference between the alternative hypotheses of each test and the corresponding rejection regions. The omnibus test is directed toward any point θ in the parameter space \mathbf{R}^K away from the origin. In many analyses, however, it may be of scientific interest to use a test directed toward a restricted alternative hypothesis represented by a subregion of \mathbf{R}^K . In this case, the omnibus test will not be as powerful as a test of such a restricted alternative, when that alternative is true.

The Mantel-Haenszel test is often called a test of no partial association, or just association, because it is directed to such a restricted alternative hypothesis. To show how this test relates to the omnibus test, and as a prelude to developing other tests of the joint null hypothesis H_0 in (4.48), it is instructive to show that the null hypothesis can be partitioned into the intersection of two subhypotheses as follows:

<i>Omnibus</i>		<i>Homogeneity</i>		<i>No Partial Association</i>
$H_0: \boldsymbol{\theta} = \mathbf{J}\theta_0$ or $\theta_k = \theta_0 \forall k$	\equiv	$H_{0H}: \theta_k = \theta \quad \forall k$ and		$H_{0A}: \theta = \theta_0$
$K \, df$	$=$	$K - 1 \, df$	$+$	$1 \, df$
$H_1: \boldsymbol{\theta} \neq \mathbf{J}\theta_0$	\equiv	$H_{1H}: \theta_k \neq \theta_\ell, \quad k \neq \ell$, or		$H_{1A}: \theta_k = \theta \neq \theta_0 \forall k$

(4.55)

The omnibus joint null hypothesis H_0 in (4.48) states that the K components of $\boldsymbol{\theta}$ are all equal to θ_0 , and thus the test has K degrees of freedom. The null hypothesis of *homogeneity* H_{0H} specifies that the components of $\boldsymbol{\theta}$ share a common value θ , possibly $\neq \theta_0$. This, in turn, implies that we can define $K - 1$ independent contrasts among the values of $\boldsymbol{\theta}$, each of which has expectation zero under H_{0H} , and thus this test has $K - 1 \, df$. The null hypothesis of no *partial association* H_{0A} specifies that the

assumed common value for all the components of θ equals the null hypothesis value. Clearly, $H_0 \equiv H_{0H} \cap H_{0A}$. In terms of a linear model, such as a simple ANOVA, the hypothesis of homogeneity H_{0H} corresponds to the hypothesis of no group by stratum interaction effect, and the hypothesis of no association H_{0A} corresponds to the hypothesis of no overall group effect.

For example, for $\theta = \log$ (odds ratio), the null hypothesis for the Mantel-Haenszel test in (4.13) can be formally expressed as

$$\begin{aligned} H_{0(MH)}: & (\theta_1 = \dots = \theta_K = \theta = \theta_0 = 0) \\ & \equiv (\theta_1 = \dots = \theta_K = \theta) \cap (\theta = \theta_0 = 0) \\ & \equiv H_{0H} \cap H_{0A}. \end{aligned} \quad (4.56)$$

Therefore, the Mantel-Haenszel test assumes that the hypothesis H_{0H} of homogeneity is true, in which case the test is maximally efficient (as we shall see shortly).

The general alternative hypothesis, that the components do not equal the null hypothesis value for all strata, then is the union of the alternative hypothesis of some heterogeneity among the components, or the assumed common component being different from the null value. If there is heterogeneity of the odds ratios in the population, meaning that $\theta_k \neq \theta_\ell$ for some $k \neq \ell$ so that H_{0H} is false, then the Mantel-Haenszel test might be viewed as testing the hypothesis that the *average* of the stratum-specific log odds ratios is zero, or solely testing the hypothesis H_{0A} : $\theta = \theta_0 = 0$. In this case, however, the lack of homogeneity in the stratum-specific parameters will reduce the power of the stratified-adjusted Mantel-Haenszel test as a test of H_{0A} alone. In fact, when H_{0H} is false and there is heterogeneity of the odds ratios over strata in the population, the Mantel-Haenszel test may no longer be appropriate. Alternative models in this case are described in Section 4.10.

4.6 TESTS OF HOMOGENEITY

A variety of methods are available to test the hypothesis of homogeneity among the measures of association $\{\theta_k\}$ on some scale $\theta_k = G(\pi_{1k}, \pi_{2k})$:

$$H_{0H}: \theta_1 = \theta_2 = \dots = \theta_K, \quad (4.57)$$

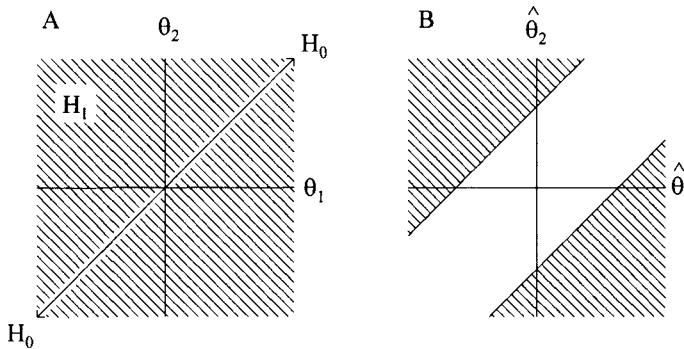
against the alternative that there is a difference between at least two strata

$$H_{1H}: \theta_k \neq \theta_\ell \text{ some } 1 \leq k < \ell \leq K. \quad (4.58)$$

Note that in this case, it is irrelevant whether $\theta_k = \theta_0$ for any or all strata.

For the case of only two strata, these hypotheses are depicted in Figure 4.3A. The null hypothesis is that the parameter values correspond to one of the possible points that lie on the line of equality. The alternative hypothesis is that the values lie somewhere in the two-dimensional space away from this line of equality.

Fig. 4.3 The null (H_0) and alternative (H_1) hypotheses for the test of homogeneity with two parameters $\theta = (\theta_1, \theta_2)$ (A) and the corresponding rejection region (B).



4.6.1 Contrast Test of Homogeneity

The null hypothesis of homogeneity H_{0H} implies that the difference between any two strata is zero, $\theta_k - \theta_\ell = 0$ for all $k \neq \ell$. Thus, it is possible to define $K - 1$ contrasts among the sample estimates using a $K \times (K - 1)$ contrast matrix \mathbf{C} of rank $K - 1$ to test the hypothesis H_{0C} : $\mathbf{C}'\theta = 0$, where $\mathbf{C}'\theta$ is a $(K - 1) \times 1$ vector. A variety of such contrast matrices could be defined. The contrast matrix of successive differences $\{\theta_k - \theta_{k+1}\}$ for $k = 1, \dots, (K - 1)$ is of the form

$$\mathbf{C}' = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}, \quad (4.59)$$

whereas in the matrix of contrasts of each stratum versus the last $\{\theta_k - \theta_K\}$ the -1 in each row is moved to the last column. The contrast matrix of each stratum versus the simple average value over all strata $\{\theta_k - \bar{\theta}\}$ is of the form

$$\mathbf{C}' = \frac{1}{K} \begin{bmatrix} K - 1 & -1 & \cdots & -1 & -1 \\ -1 & K - 1 & \cdots & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & \cdots & K - 1 & -1 \end{bmatrix}, \quad (4.60)$$

where

$$\theta_k - \theta = \theta_k - \frac{\sum_\ell \theta_\ell}{K} = \frac{K\theta_k - \sum_\ell \theta_\ell}{K} \propto (K-1)\theta_k - \sum_{\ell \neq k} \theta_\ell. \quad (4.61)$$

For any such contrast matrix

$$H_{0H} \Leftrightarrow H_{0C}: \mathbf{C}'\boldsymbol{\theta} = \mathbf{0} \quad (4.62)$$

$$H_{1H} \Leftrightarrow H_{1C}: \mathbf{C}'\boldsymbol{\theta} \neq \mathbf{0}.$$

Since $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed as in (4.46), then the test of homogeneity is provided by the T^2 -like Wald statistic, defined as the quadratic form

$$X_H^2 = (\mathbf{C}'\hat{\boldsymbol{\theta}})' (\mathbf{C}'\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} \mathbf{C})^{-1} \mathbf{C}'\hat{\boldsymbol{\theta}}, \quad (4.63)$$

which is asymptotically distributed as χ_{K-1}^2 on $K-1$ *df*. Note that the test is computed using the estimate of the covariance matrix defined under the general hypothesis of homogeneity, with no restrictions that the parameter values equal θ_0 , unlike the omnibus test statistic presented in (4.50).

The value of the test statistic is the same for all such contrast matrices that are of rank $K-1$ and that satisfy $\mathbf{C}'\mathbf{J} = \mathbf{0}$ under H_{0H} , where \mathbf{J} is the unit vector (cf. Anderson, 1984, p. 170). This test can be computed directly using a custom routine programmed in a matrix language such as SAS PROC IML, or by using the SAS procedure PROC CATMOD with an appropriate response function. PROC CATMOD uses the method of weighted least squares described by Grizzle et al. (1969) to fit models describing the response probabilities associated with covariate values for each subpopulation (stratum). The test obtained from the noniterative weighted least squares fit is algebraically equivalent to that above for the same response function. For example, if $\boldsymbol{\theta}$ is defined as the log relative risk, then the equivalent response function in the GSK procedure (PROC CATMOD) is the log of the probability. The test results then are identical.

For the case of two correlated measures such as those described in Figure 4.2.B, the test statistic X_H^2 in (4.63) using the successive difference contrast matrix (4.59) reduces to $X_H^2 = (\hat{\theta}_1 - \hat{\theta}_2)^2 / \hat{V}(\hat{\theta}_1 - \hat{\theta}_2)$, where $\hat{V}(\hat{\theta}_1 - \hat{\theta}_2) = \hat{\sigma}_{\hat{\theta}_1}^2 + \hat{\sigma}_{\hat{\theta}_2}^2 - 2\widehat{Cov}(\hat{\theta}_1, \hat{\theta}_2)$. Thus, the null hypothesis reduces to $H_{0H}: \theta_1 = \theta_2$, in which case the rejection region boundary is defined as all points $(\hat{\theta}_1, \hat{\theta}_2)$ for which $X_H^2 = X_{1(1-\alpha)}^2$ or for which $|\hat{\theta}_1 - \hat{\theta}_2| = \sqrt{3.841 \times \hat{V}(\hat{\theta}_1 - \hat{\theta}_2)}$, 3.841 being the 95th percentile of the 1 *df* chi-square distribution. For the bivariate example, the rejection region is defined by the parallel lines shown in Figure 4.3B for which the observations leading to rejection of H_{0H} principally fall in quadrants II and IV, wherein the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ differ in sign.

4.6.2 Cochran's Test of Homogeneity

Cochran (1954b) proposed a test of homogeneity for odds ratios based on the sums of squares of deviations of the stratum-specific odds ratios about the mean odds ratio. This idea can be generalized to a test for homogeneity on any scale. To obtain the *MVLE* of the assumed common value for measures of association $\{\theta_k\}$ on the scale $\theta_k = G(\pi_{1k}, \pi_{2k})$ we used weights inversely proportional to the variances as described in Section 4.3.5. Under the hypothesis of homogeneity H_{0H} in (4.57), the stratum-specific estimate for the k th stratum is asymptotically distributed as

$$\hat{\theta}_k - \theta \xrightarrow{d} N\left(0, \sigma_{\hat{\theta}_k}^2\right) \quad (4.64)$$

for $1 \leq k \leq K$. Since the hypothesis of homogeneity does not require that $\pi_{1k} = \pi_{2k}$, the variance of $\hat{\theta}_k$ is evaluated assuming that $\pi_{1k} \neq \pi_{2k}$, as described in Section 2.3.

For now assume that the variances, and thus their inverse ($\tau_k = \sigma_{\hat{\theta}_k}^{-2}$) are known. From (4.37), since $\hat{\theta} \xrightarrow{p} \theta$, then asymptotically, from Slutsky's theorem (A.44),

$$\sqrt{\tau_k} (\hat{\theta}_k - \hat{\theta}) \xrightarrow{p} \sqrt{\tau_k} (\hat{\theta}_k - \theta) \xrightarrow{d} \mathcal{N}(0, 1). \quad (4.65)$$

Further, since $\hat{\sigma}_{\hat{\theta}_k}^2 \xrightarrow{p} \sigma_{\hat{\theta}_k}^2$ and $\hat{\tau}_k \xrightarrow{p} \tau_k$, then

$$\hat{\tau}_k (\hat{\theta}_k - \hat{\theta})^2 \xrightarrow{d} \chi_1^2. \quad (4.66)$$

Therefore,

$$X_{H,C}^2 = \sum_k \hat{\tau}_k (\hat{\theta}_k - \hat{\theta})^2 \quad (4.67)$$

is distributed asymptotically as chi-square on $K - 1$ *df*; $K - 1$ because we estimate θ as a linear combination of the K stratum-specific estimates. Algebraically, Cochran's test is equivalent to the contrast test X_H^2 in (4.63), which, in turn, is equivalent to the GSK test obtained from the SAS PROC CATMOD.

Example 4.10 Clinical Trial in Duodenal Ulcers (continued)

For the ulcer drug clinical trial example, Cochran's test of homogeneity for the odds ratios is based on the elements

Stratum	$\hat{\theta}_k = \log(\widehat{OR}_k)$	$\hat{\tau}_k = \hat{\sigma}_{\hat{\theta}_k}^{-2}$
1	-0.1854	5.31919
2	1.3218	1.11801
3	1.0015	5.27749

and $X_{H,C}^2 = 4.5803$ on 2 df with $p \leq 0.102$. The tests of homogeneity for the three principal measures of association, each on 2 df , are

Measure	$X_{H,C}^2$	$p \leq$
Risk differences	4.797	0.091
log Relative risks	3.648	0.162
log Odds ratios	4.580	0.102

For this example, based on the relative values of the test statistics, there is the least heterogeneity among the relative risks, and the most heterogeneity among the risk differences. However, the difference in the extent of heterogeneity among the three scales is slight.

Example 4.11 *Religion and Mortality (continued)*

For the study of religion versus mortality, the tests of homogeneity on 3 df for each of the three scales likewise are

Measure	$X_{H,C}^2$	$p \leq$
Risk differences	3.990	0.263
Relative risks	0.929	0.819
Odds ratios	1.304	0.728

In this example, there is some heterogeneity among the risk differences, but far from significant, and almost none among the relative risks or odds ratios.

4.6.3 Zelen's Test

A computationally simple test of homogeneity was proposed by Zelen (1971). Let X_A^2 designate a test of the hypothesis H_{0A} of no partial association (or just association) for the average measure of association among the K strata, such as the Cochran-Mantel-Haenszel test on 1 df of the common odds ratio. Since the omnibus null hypothesis can be partitioned as shown in (4.55), Zelen (1971) proposed that the test of homogeneity

$$X_{H,Z}^2 = X_O^2 - X_A^2 \quad (4.68)$$

be obtained as the difference between the omnibus chi-square test X_O^2 on K df , and the test of association X_A^2 on 1 df . For the ulcer clinical trial example using the conditional Mantel-Haenszel test yields $X_{H,Z}^2 = 7.6239 - 3.00452 = 4.62$, which is slightly greater than the Cochran test value $X_{H,C}^2 = 4.58$.

Mantel et al. (1977) and Halperin et al. (1977) criticize this test and present examples that show that this simple test may perform poorly in some situations. The problem is that an optimal test of the null hypothesis of homogeneity H_{0H} should use the variances estimated under that hypothesis. Thus, both the contrast test X_H^2 in (4.63) and Cochran's test $X_{H,C}^2$ in (4.67) use the variances $\widehat{\Sigma}_{\widehat{\theta}}$ estimated under the general alternative hypothesis H_{1O} in (4.49) that some of the $\{\theta_k\}$, if not all, differ from the null value θ_0 . However, both X_A^2 and X_O^2 use the variances defined

under the general null hypothesis H_0 , Σ_0 , so that equality does not hold, that is, $X_O^2 \neq X_H^2 + X_A^2$. In general, therefore, this test should be avoided.

4.6.4 Breslow-Day Test for Odds Ratios

Breslow and Day (1980) also suggested a test of homogeneity for odds ratios for use with a Mantel-Haenszel test that is based on the Mantel-Haenszel estimate of the common odds ratio \widehat{OR}_{MH} in (4.14). In the k th stratum, given the margins for that 2×2 table $(m_{1k}, m_{2k}, n_{1k}, n_{2k})$, then the expectation of the index frequency a_k under the hypothesis of homogeneity $OR_k = OR$ can be estimated as

$$\widehat{E}(a_k | \widehat{OR}_{MH}) = \tilde{a}_k \text{ such that } OR_k = \widehat{OR}_{MH}. \quad (4.69)$$

This expected frequency is the solution to

$$\frac{(\tilde{a}_k)(n_{2k} - m_{1k} + \tilde{a}_k)}{(m_{1k} - \tilde{a}_k)(n_{1k} - \tilde{a}_k)} = \widehat{OR}_{MH}, \quad (4.70)$$

(see also Problem 2.7.2). Solving for \tilde{a}_k yields

$$\begin{aligned} m_{1k}n_{1k}\widehat{OR}_{MH} = & \tilde{a}_k \left[n_{2k} - m_{1k} + \widehat{OR}_{MH}(n_{1k} + m_{1k}) \right] \\ & + \tilde{a}_k^2 \left(1 - \widehat{OR}_{MH} \right), \end{aligned} \quad (4.71)$$

which is a quadratic function in \tilde{a}_k . The root such that $0 < \tilde{a}_k \leq \min(n_{1k}, m_{1k})$ yields the desired estimate. Given the margins of the table $(n_{1k}, n_{2k}, m_{1k}, m_{2k})$ the expected values of the other cells of the table are obtained by subtraction, such as $\tilde{b}_k = m_{1k} - \tilde{a}_k$.

Then the Breslow-Day test of homogeneity of odds ratios is a Pearson contingency test of the form

$$X_{H,BD}^2 = \sum_{k=1}^K \left[\frac{(a_k - \tilde{a}_k)^2}{\tilde{a}_k} + \frac{(b_k - \tilde{b}_k)^2}{\tilde{b}_k} + \frac{(c_k - \tilde{c}_k)^2}{\tilde{c}_k} + \frac{(d_k - \tilde{d}_k)^2}{\tilde{d}_k} \right]. \quad (4.72)$$

Since the term in the numerator is the same for each cell of the k th stratum, for example, $(b_k - \tilde{b}_k)^2 = (a_k - \tilde{a}_k)^2$, this statistic can be expressed as

$$X_{H,BD}^2 = \sum_{k=1}^K \frac{(a_k - \tilde{a}_k)^2}{\widehat{V}(a_k | \widehat{OR}_{MH})}, \quad (4.73)$$

where

$$\widehat{V}(a_k | \widehat{OR}_{MH}) = \left[\frac{1}{\tilde{a}_k} + \frac{1}{\tilde{b}_k} + \frac{1}{\tilde{c}_k} + \frac{1}{\tilde{d}_k} \right]^{-1}. \quad (4.74)$$

This test for homogeneity of odds ratios is used in SAS PROC FREQ as part of the Cochran-Mantel-Haenszel analysis with the *cmh* option. For the data from Example 4.1, this test yields the value $X_{H,BD}^2 = 4.626$ on 2 *df* with $p \leq 0.099$, and for Example 4.6, this test value is $X_{H,BD}^2 = 1.324$ on 3 *df* with $p \leq 0.7234$. In both cases, the test value is slightly larger than the Cochran test value, the *p*-values smaller.

4.6.5 Tarone Test for Odds Ratios

Breslow and Day (1980) suggested that $X_{H,BD}^2$ is distributed asymptotically as χ_{K-1}^2 on $K-1$ *df*. Tarone (1985) showed that this would be the case if a fully efficient estimate of the common odds ratio, such as the *MLE*, were used as the basis for the test. Since the Mantel-Haenszel estimate is not fully efficient, then $X_{H,BD}^2$ is stochastically larger than a variate distributed as χ_{K-1}^2 . Tarone also showed that a corrected test can be obtained as

$$X_{H,BD,T}^2 = X_{H,BD}^2 - \frac{\left(\sum_{k=1}^K a_k - \sum_{k=1}^K \tilde{a}_k\right)^2}{\sum_{k=1}^K \widehat{V}(a_k | \widehat{OR}_{MH})}, \quad (4.75)$$

that is asymptotically distributed as χ_{K-1}^2 on $K-1$ *df*. Breslow (1996) recommends that in general the corrected test should be preferred to the original Breslow-Day test, but also points out that the correction term is often negligible. For the data in Example 4.1, the corrected test value $X_{H,BD,T}^2 = 4.625$ on 2 *df* with $p \leq 0.100$, and for Example 4.6, $X_{H,BD,T}^2 = 1.3236$ on 3 *df* with $p \leq 0.7235$: in both cases, nearly identical to the original Breslow-Day test.

The Tarone modification of the test is also provided by SAS PROC FREQ using the *bdt* tables option.

4.7 EFFICIENT TESTS OF NO PARTIAL ASSOCIATION

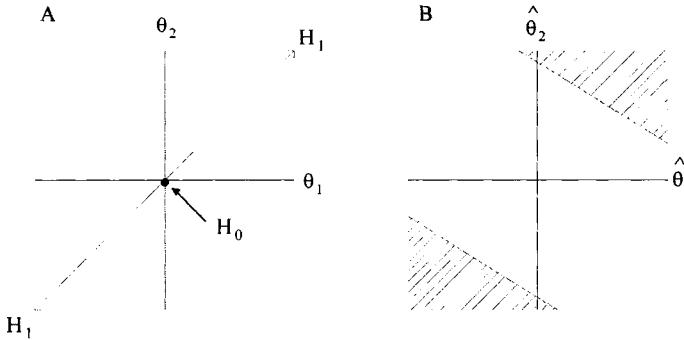
From the partitioning of hypotheses presented in (4.55), the global null hypothesis in (4.48) has two components: that the measures $\{\theta_k\}$ among strata are all equal $\{\theta_k = \theta\}$ (homogeneity) and that they are all equal to the null value $\{\theta = \theta_0\}$ (association). Clearly, if the hypothesis of homogeneity is rejected, then so is the global null hypothesis because if $\theta_k \neq \theta_\ell$ for any two strata k and ℓ , then both cannot also equal θ_0 . However, the hypothesis of homogeneity could be satisfied when all strata have a common measure that differs from the null value. Thus, we desire an efficient test that is directed specifically toward the association hypothesis under the assumption of homogeneity.

4.7.1 Restricted Alternative Hypothesis of Association

The principal disadvantage of omnibus T^2 -like test of the joint null hypothesis for the K strata is that it is directed to the global alternative hypothesis that a difference of some magnitude in either direction exists for at least one strata. Usually, however, one is interested in declaring that an overall difference of some magnitude exists over all strata, or on average in some sense in the overall population. Thus, a statistically significant omnibus test may not be scientifically relevant or clinically compelling.

Alternatively, a test directed against the restricted alternative hypothesis of a common difference on some scale for all strata has more scientific appeal, and will be more powerful than the omnibus test to detect a common difference when such is

Fig. 4.4 The null (H_0) and alternative (H_1) hypotheses for the test of association with two parameters $\theta = (\theta_1, \theta_2)$ (A) and the corresponding rejection region (B).



the case. The null and alternative hypotheses in this case are

$$\begin{aligned} H_{0A} : \theta_1 &= \theta_2 = \dots = \theta_K = \theta = \theta_0 & (4.76) \\ H_{1A} : \theta_1 &= \theta_2 = \dots = \theta_K = \theta \neq \theta_0. \end{aligned}$$

The null hypothesis is the same as that for the K df omnibus test; however, this test has a restricted alternative hypothesis that a constant difference $\theta \neq \theta_0$ exists on some scale. Thus, $H_{1A} \subset H_{1O}$ for the omnibus test.

For the case of a bivariate analysis as described in Figure 4.2 for the omnibus test, Figure 4.4A depicts the null and alternative hypotheses for the test of association. As for the omnibus test, the null hypothesis corresponds to the origin in the two-dimensional parameter space. The alternative hypothesis, however, consists of all points falling on the positive and negative projections corresponding to the line of equality $\theta_1 = \theta_2$. Points in the parameter space for which $\theta_1 \neq \theta_2$ are not of interest. Thus, when H_{1A} is true for θ defined on some scale $G(\pi_1, \pi_2)$, this test will have greater power than the K df omnibus test.

For illustration, as shown in (4.20), the Mantel-Haenszel test can be viewed as an estimation-based test using an estimate of the common value of the parameter $\hat{\theta}$ and an estimate of its variance under H_0 , $\hat{\sigma}_{\hat{\theta}|H_0}^2$. In such a construction, $\hat{\theta}$ would be a linear function of the $\{\hat{\theta}_k\}$ and the test could be expressed as

$$Z_A = \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}|H_0}} = \sum_k a_k \hat{\theta}_k, \quad (4.77)$$

where asymptotically Z_A is distributed as $N(0, 1)$. As with the MVLE, the estimate $\hat{\theta}$ would be obtained as the weighted average of the $\{\theta_k\}$ using weights $\{a_k\}$ inversely proportional to the variances, in which case the variances would be estimated under the null hypothesis $\{\hat{\sigma}_{0k}^2\}$. Thus, the elements of $\{\hat{\theta}_k\}$ with smaller variance would receive greater weight.

For the bivariate example in Figure 4.2B, where the two sample statistics are correlated, a similar statistic $Z_A = a_1\hat{\theta}_1 + a_2\hat{\theta}_2$ is employed. (Because of the correlation, the actual weights are slightly different from those with independent strata.) Setting $Z_A = \pm Z_{1-\alpha/2} = \pm 1.96$ for $\alpha = 0.05$, we can solve for the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ that satisfy this equality to determine the lines defining the upper and lower rejection regions. Figure 4.4B shows the corresponding rejection regions for this bivariate example. These rejection regions contain observations principally in Quadrants I and III, but also admit cases in quadrants II and IV where either statistic $\hat{\theta}_1$ or $\hat{\theta}_2$ is quite distant from zero. When the bivariate statistics $\hat{\theta}_1$ and $\hat{\theta}_2$ have equal variances and are assigned equal weight, then the lines defining the rejection region lie perpendicular to the projection of the alternative hypothesis. However, when the weights differ, as in this example, then the line of rejection is tilted toward the origin for the sample statistic that has the smaller variance and receives the greater weight.

4.7.2 Radhakrishna Family of Efficient Tests of Association

We now show that the Cochran and Mantel-Haenszel tests are such tests of association under the hypothesis that a common odds ratio applies to all K strata. These tests were derived rather heuristically, although Cochran does provide a justification for his test.

Radhakrishna (1965) considered the problem more generally from the perspective of deriving a family of asymptotically fully efficient tests of H_{0A} versus H_{1A} for multiple 2×2 tables with measures of association $\{\theta_k\}$ on some scale $\theta = g(\pi_1) - g(\pi_2)$ such that $H_0: \pi_{1k} = \pi_{2k} = \pi_k$ implies, and is implied by

$$H_{0A(g)}: g(\pi_{1k}) - g(\pi_{2k}) = \theta_0 = 0. \quad (4.78)$$

Specifically, we desire an asymptotically efficient test of H_0 against a restricted alternative hypothesis of the form

$$H_{1A(g)}: g(\pi_{1k}) - g(\pi_{2k}) = \theta \text{ for } \forall k, \quad (4.79)$$

or that there is a constant difference θ on the specified scale $g(\pi)$. This family includes a test of a common log odds ratio for which $g(\pi) = \log[\pi/(1 - \pi)]$ (the logit), a test of a common log relative risk for which $g(\pi) = \log(\pi)$, and a test of a common risk difference for which $g(\pi) = \pi$ (the identity function), among others. Radhakrishna's initial intent was to explore the asymptotic efficiency of Cochran's stratified-adjusted test, X_U^2 , but in doing so he described a much broader family of tests.

In Problem 4.1 it is readily shown that the Cochran test of association in (4.12), using the unconditional variance, can be expressed as

$$X_{U,C}^2 = \frac{[\sum_k \hat{w}_k (p_{1k} - p_{2k})]^2}{\sum_k \hat{w}_k^2 \hat{V}_{0k}}, \quad (4.80)$$

with $\hat{V}_{0k} = \hat{\sigma}_{0k}^2 = \hat{V}(p_{1k} - p_{2k} | H_0)$ as presented in (4.53) and where

$$\hat{w}_k = \left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right)^{-1} = \frac{n_{1k} n_{2k}}{N_k}. \quad (4.81)$$

It is not obvious, however, that these weights are optimal in any statistical sense. Note that these weights $\{\hat{w}_k\}$ differ from the weights $\{\omega_k\}$ used earlier to obtain the MVLE of the common log odds ratio.

Radhakrishna (1965) proposed that the weights be derived so as to maximize the asymptotic efficiency of the test. To a first-order Taylor's approximation about an intermediate value $\pi_k \in (\pi_{1k}, \pi_{2k})$ in the k th stratum under a local alternative it can be shown (see Problem 4.4) that asymptotically

$$\theta_k \cong g'(\pi_k)(\pi_{1k} - \pi_{2k}), \quad (4.82)$$

so that

$$\pi_{1k} - \pi_{2k} \cong \theta_k / g'(\pi_k). \quad (4.83)$$

Under H_{1A} : $\theta_k = \theta$ for all strata, this implies that a test based on a weighted average of the $\{p_{1k} - p_{2k}\}$ should be asymptotically fully efficient to test H_{0A} versus H_{1A} above.

For a measure of association θ corresponding to a given scale with function $g(\pi)$, we then desire an asymptotically most powerful or fully efficient test statistic of the form

$$T = \sum_k w_k (p_{1k} - p_{2k}). \quad (4.84)$$

The asymptotic variance of the statistic evaluated under the null hypothesis is $V(T|\theta_0) = \sum_k w_k^2 \sigma_{0k}^2$, where $\sigma_{0k}^2 = V[(p_{1k} - p_{2k}) | H_0]$, as defined in (4.52).

From (3.78) the Pitman efficiency of such a test statistic is

$$Eff(T) = \left[\left(\frac{dE(T)}{d\theta} \right)^2 \bigg/ V(T|\theta) \right]_{\theta=\theta_0}. \quad (4.85)$$

Thus, we seek the expression for the weights $\{w_k\}$ that will maximize the efficiency of the test for a given scale $g(\pi)$. From the first-order approximation in (4.83) under the restricted alternative hypothesis that $\theta_k = \theta \forall k$, it follows that asymptotically the expected value of the statistic is

$$E(T|\theta) = \sum_k w_k (\pi_{1k} - \pi_{2k}) \cong \sum_k \frac{w_k \theta_k}{g'(\pi_k)} = \theta \sum_k \frac{w_k}{g'(\pi_k)}. \quad (4.86)$$

Therefore,

$$\frac{dE(T)}{d\theta} = \sum_k w_k g'(\pi_k)^{-1}, \quad (4.87)$$

and

$$Eff(T) = \frac{\left[\sum_k w_k g'(\pi_k)^{-1} \right]^2}{\sum_k w_k^2 \sigma_{0k}^2}. \quad (4.88)$$

To obtain the set of weights for which the $Eff(T)$ is maximized it is convenient to express this result in matrix terms. Let $\mathbf{W} = (w_1 \cdots w_K)^T$, $\mathbf{G} = [g'(\pi_1)^{-1} \cdots g'(\pi_K)^{-1}]^T$, and $\Sigma_0 = \text{diag}(\sigma_{01}^2 \cdots \sigma_{0K}^2)$. Then

$$Eff(T) = \frac{(\mathbf{W}^T \mathbf{G})^2}{\mathbf{W}^T \Sigma_0 \mathbf{W}}. \quad (4.89)$$

The value of w for which the $Eff(T)$ is maximized can be obtained from a basic theorem on the extrema of quadratic forms that follows from the *Cauchy-Schwartz Inequality*: $(x^T y)^2 \leq (x^T x)(y^T y)$. Let $\Sigma_0^{1/2}$ be the root matrix to Σ_0 defined such that $\Sigma_0^{1/2} \Sigma_0^{1/2} = \Sigma_0$ and $\Sigma_0^{-1/2} \Sigma_0 \Sigma_0^{-1/2} = I_K$, the identity matrix of order K . Then from the Cauchy-Schwartz inequality,

$$(\mathbf{W}^T \mathbf{G})^2 = (\mathbf{W}^T \Sigma_0^{1/2} \Sigma_0^{-1/2} \mathbf{G})^2 \leq (\mathbf{W}^T \Sigma_0 \mathbf{W}) (\mathbf{G}^T \Sigma_0^{-1} \mathbf{G}), \quad (4.90)$$

and

$$\frac{(\mathbf{W}^T \mathbf{G})^2}{\mathbf{W}^T \Sigma_0 \mathbf{W}} \leq (\mathbf{G}^T \Sigma_0^{-1} \mathbf{G}). \quad (4.91)$$

The equality, and thus the maxima, is obtained using a vector proportional to $\mathbf{W} = \Sigma_0^{-1} \mathbf{G}$ that yields

$$\frac{(\mathbf{W}^T \mathbf{G})^2}{\mathbf{W}^T \Sigma_0 \mathbf{W}} = \frac{(\mathbf{G}^T \Sigma_0^{-1} \mathbf{G})^2}{(\mathbf{G}^T \Sigma_0^{-1} \Sigma_0 \Sigma_0^{-1} \mathbf{G})} = \frac{(\mathbf{G}^T \Sigma_0^{-1} \mathbf{G})^2}{(\mathbf{G}^T \Sigma_0^{-1} \mathbf{G})} = \mathbf{G}^T \Sigma_0^{-1} \mathbf{G}. \quad (4.92)$$

Thus, we have the well-known result (cf. Rao, 1973, p. 60)

$$\max_{\mathbf{W}} \frac{(\mathbf{W}^T \mathbf{G})^2}{\mathbf{W}^T \Sigma_0 \mathbf{W}} = \mathbf{G}^T \Sigma_0^{-1} \mathbf{G}, \quad (4.93)$$

with the maximum obtained using a vector of weights $\mathbf{W} \propto \Sigma_0^{-1} \mathbf{G}$. Note that the weights need not sum to unity because multiplying by any norming constant will cancel from the numerator and denominator.

For the case of K independent strata, Σ_0 is a diagonal matrix so that the vector of weights which maximizes $Eff(T)$ is

$$\mathbf{W} = \Sigma_0^{-1} \mathbf{G} = \left[\frac{g'(\pi_1)^{-1}}{\sigma_{01}^2} \cdots \frac{g'(\pi_K)^{-1}}{\sigma_{0K}^2} \right]^T. \quad (4.94)$$

Table 4.8 Radhakrishna test weights for each scale of association.

Scale	Derivative	Optimal Weight
$\hat{\theta}_k = g(p_{1k}) - g(p_{2k})$	$g'(p_k)$	$\hat{w}_k = [\hat{V}_{0k} g'(p_k)]^{-1}$
<i>Risk Difference</i>		
$p_{1k} - p_{2k}$	1.0	$\hat{V}_{0k}^{-1} = \frac{n_{1k}n_{2k}}{N_k p_k (1 - p_k)}$
<i>log Relative Risk</i>		
$\log(p_{1k}) - \log(p_{2k})$	$\frac{1}{p_k}$	$\frac{p_k}{\hat{V}_{0k}} = \frac{n_{1k}n_{2k}}{N_k (1 - p_k)}$
<i>log Odds Ratio</i>		
$\log \frac{p_{1k}}{1 - p_{1k}} - \log \frac{p_{2k}}{1 - p_{2k}}$	$\frac{1}{p_k (1 - p_k)}$	$\frac{p_k (1 - p_k)}{\hat{V}_{0k}} = \frac{n_{1k}n_{2k}}{N_k}$

Thus, the optimal test weight for the k th stratum is

$$w_k = \frac{1}{g'(\pi_k) \sigma_{0k}^2}. \quad (4.95)$$

A consistent estimate of \mathbf{W} is provided by $\widehat{\mathbf{W}}$ with elements

$$\widehat{w}_k = [g'(p_k) \widehat{\sigma}_{0k}^2]^{-1} = \left[\frac{dg(p_k)}{dp_k} \widehat{V}_{0k} \right]^{-1}, \quad (4.96)$$

where $\widehat{V}_{0k} = \widehat{\sigma}_{0k}^2$ is as presented in (4.53). Then the test that maximizes the asymptotic efficiency for the scale $g(\pi)$ is

$$X_{A(g)}^2 = \frac{[\sum_k \widehat{w}_k (p_{1k} - p_{2k})]^2}{\sum_k \widehat{w}_k^2 \widehat{V}_{0k}} = \frac{T_g^2}{\widehat{V}(T_g | H_0)}, \quad (4.97)$$

where

$$\widehat{V}(T_g | H_0) = \sum_k \frac{\widehat{V}_{0k}}{\left[g'(p_k) \widehat{V}_{0k} \right]^2} = \sum_k \frac{1}{g'(p_k)^2 \widehat{V}_{0k}}, \quad (4.98)$$

and where $X_{A(g)}^2$ is asymptotically distributed as χ^2 on 1 *df* under H_0 . This provides an asymptotically efficient test of H_{0A} versus H_{1A} for any parameter that can be expressed as a scale transformation of the form $\theta_k = g(\pi_{1k}) - g(\pi_{2k})$.

Note that because the test employs estimated weights $\{\widehat{w}_k\}$, there is a small loss of efficiency in the finite sample case; see Bloch and Moses (1988). However, since the estimated weights $\{\widehat{w}_k\}$ are consistent estimates of the true optimal weights, then from Slutsky's theorem (A.45) the test is asymptotically fully efficient.

For the principal measures of association employed in Chapter 2, Table 4.8 summarizes the elements of the optimal weights \widehat{w}_k . The test for a specific scale can then

be obtained by substituting the corresponding weights into the weighted Cochran-Mantel-Haenszel statistic in (4.97). This leads to a specific test that is directed to the specific alternative hypothesis of association on a specific scale $g(\pi)$. The precise expressions for the tests for a common risk difference, odds ratio, and relative risk are derived in Problem 4.4.3.

Each of these tests is a weighted linear combination of the risk differences within the K strata. Therefore, any one test is distinguished from the others by the *relative* magnitudes of the weights for each strata, or $w_k/\sum_\ell w_\ell$. Also, under $H_0: \pi_{1k} = \pi_{2k}$ for all strata (k), any nonzero set of weights $\{\hat{w}_k\}$ yields a weighted test that is asymptotically distributed as χ_1^2 , since asymptotically

$$\begin{aligned} E \left[\sum_k \hat{w}_k (p_{1k} - p_{2k}) \right] &\cong \sum_k w_k (\pi_{1k} - \pi_{2k}) = 0 \\ V \left[\sum_k \hat{w}_k (p_{1k} - p_{2k}) \right] &\cong \sum_k w_k^2 \sigma_{0k}^2 \end{aligned} \quad (4.99)$$

and asymptotically $X_A^2 \stackrel{d}{\approx} \chi^2$ on 1 *df*. Therefore, the test for each scale $g(\pi)$ is of size α under H_0 . However, the tests differ in their power under specific alternatives.

Example 4.12 *Clinical Trial in Duodenal Ulcers (continued)*

For the ulcer drug clinical trial example, the differences within each stratum, the optimal weights $\{\hat{w}_k\}$ for the test of association on each scale (*RD*, $\log RR$, and $\log OR$), and the relative magnitudes of the weights for each scale expressed as a percentage of the total are

k	$p_{1k} - p_{2k}$	\hat{w}_k			<i>Proportional Weights</i>		
		<i>RD</i>	$\log RR$	$\log OR$	<i>RD</i>	$\log RR$	$\log OR$
1	-0.04458	92.08	37.25	22.18	0.45	0.39	0.46
2	0.30556	21.81	13.50	5.14	0.11	0.14	0.10
3	0.24506	90.00	44.00	22.49	0.44	0.46	0.45
<i>Total</i>		203.89	94.75	49.81	1.00	1.00	1.00

For example, for the first stratum ($k = 1$),

$$\begin{aligned} \hat{w}_{(RD)1} &= [p_1 (1 - p_1)]^{-1} \left(\frac{n_{11} n_{21}}{N_1} \right) \\ &= [0.40449 (1 - 0.40449)]^{-1} (22.18) = 92.08, \\ \hat{w}_{(RR)1} &= (1 - p_1)^{-1} \left(\frac{n_{11} n_{21}}{N_1} \right) \\ &= (1 - 0.40449)^{-1} (22.18) = 37.25, \\ \hat{w}_{(OR)1} &= \frac{n_{11} n_{21}}{N_1} = \frac{42 \times 47}{89} = 22.18. \end{aligned}$$

Because the relative values of the weights distinguish each test from the others, in this example the tests differ principally with respect to the weights for the first two strata.

The terms that enter into the numerator and denominator of the test statistic for each scale are

Stratum	$\widehat{w}_k (p_{1k} - p_{2k})$			$\widehat{w}_k^2 \widehat{V}_{0k}$		
	RD	$\log RR$	$\log OR$	RD	$\log RR$	$\log OR$
1	-4.1048	-1.6604	-0.9888	92.0786	15.0655	5.3426
2	6.6635	4.1250	1.5714	21.8077	8.3571	1.2128
3	22.0553	10.7826	5.5111	90.0000	21.5111	5.6195
Total	24.6140	13.2472	6.0938	203.8863	44.9338	12.1749

and the resulting test statistics are

Scale	T	$\widehat{V}(T)$	X_A^2	$p \leq$
Risk difference	24.6140	203.8863	2.9715	0.085
log Relative risk	13.2472	44.9338	3.9055	0.049
log Odds ratio	6.0938	12.1749	3.0501	0.081

The test designed to detect a common log relative risk different from zero is nominally statistically significant at $p \leq 0.05$, whereas the tests designed to detect a common risk difference or a common log odds ratio are not significant. Since the tests of homogeneity for the three strata indicated the least heterogeneity among the relative risks, then we expect the test statistic for relative risks to be greater than that for the other scales.

Example 4.13 Religion and Mortality (continued)

For the assessment of the association between religious versus nonreligious and mortality, the tests of association for each of the three scales are

Scale	X_A^2	$p \leq$
Risk difference	1.2318	0.267
log Relative risk	4.1154	0.043
log Odds ratio	3.7359	0.053

In this case, the results are strongly dependent on the extent of heterogeneity among the four strata on each scale. The risk differences show the most heterogeneity, and even though the test of homogeneity is not significant, the test of association based on the risk differences is substantially less than that for the relative risks or odds ratios that showed less heterogeneity.

4.8 ASYMPTOTIC RELATIVE EFFICIENCY OF COMPETING TESTS

4.8.1 Family of Tests

The preceding examples show that the efficiency of the test of association differs from one scale to the next and that the results of the test on each scale depend on the

extent of heterogeneity over the various strata. The fundamental problem is that we now have a family \mathcal{G} of tests, each directed to a different alternative hypothesis of a constant difference on a specified scale $g(\pi)$. However, the one test that is in fact most powerful unfortunately is not known *a priori*. This will be that test directed to the one scale $g(\pi)$ for which the corresponding measures of association $\{\theta_{gk}\}$ are indeed constant over strata or nearly so. Although tempting, it is cheating to conduct the test for all members of the family (three tests herein) and then select that one for which the p -value is smallest. Clearly, if one adopts this strategy, the Type I error probability of the "test" will be increased beyond the desired size α . Thus, the test of association for a given scale will only have a test size equal to the desired level α when that scale is specified *a priori*, meaning that the specific restricted alternative to which the test is directed is prespecified *a priori*.

The same problem exists if one first conducts the test of homogeneity for all three scales and then selects the test of association for whichever scale is least significant. In fact, the two strategies are equivalent because the scale for which the test of homogeneity has the maximum p -value will also be the scale for which the test of association has the minimum p -value. This is an instance of the problem of two-stage inference (cf. Bancroft, 1972). Under the joint null hypothesis in (4.47), Lachin and Wei (1988) show that the tests of homogeneity and of association are uncorrelated so that one could partition the total Type I error probability for the two successive tests. This approach, however, is not very appealing because the size of the test of association must be some quantity less than α , thus diminishing power. Also, the specific scale to be tested must be prespecified.

Another approach to this problem is to first consider the possible loss in efficiency associated with choosing *a priori* a test on the "wrong" scale, or one for which there is more heterogeneity relative to another scale for which the corresponding test will have greater efficiency. This can be described through the assessment of the asymptotic relative efficiency (ARE) of two tests designed to detect a common difference among strata on different scales of association.

The Radhakrishna family can be described as a family of tests of H_0 directed toward a family of alternative hypotheses of the form $H_{1A(g)}$: $g(\pi_{1k}) - g(\pi_{2k}) = \theta_g$ as presented in (4.79), where $g(\pi) \in \mathcal{G}$ is a family of scales for which is $\theta_0 = 0$ under the null hypothesis H_0 . One property of the family \mathcal{G} is that if the $\{\pi_{2k}\}$ vary over strata, then strict homogeneity on one scale implies heterogeneity on another scale. For example, when $\theta_{sk} = \theta_s \forall k$ for scale g_s , if $\pi_{2k} \neq \pi_{2\ell}$ for two of the strata, say $k \neq \ell$, then for any other scale g_r , $\theta_{rk} \neq \theta_{r\ell}$ for the same pair of strata. In such cases, whichever scale provides strict homogeneity over strata, or the scale with the greatest homogeneity over strata, will provide the more powerful test.

Example 4.14 Two Homogeneous Strata

For example, consider the following case of only two strata with probabilities such that there is a common odds ratio ($g_s = \text{logit}$), and as a result, heterogeneity in the

risk differences ($g_r = \text{identity}$):

Stratum	π_{1k}	π_{2k}	$g_s = \text{logit}$		$g_r = \text{identity}$
			$\theta_{sk} = \log(OR)$	OR	$\theta_{rk} = RD$
1	0.20	0.091	0.9163	2.50	0.109
2	0.30	0.146	0.9163	2.50	0.154

In this case, the test based on the logit scale, or assuming a common odds ratio, provides the most powerful test of H_0 . Conversely, the test assuming a constant risk difference will provide less power as a test of H_0 .

4.8.2 Asymptotic Relative Efficiency

The loss in efficiency incurred when we use a test for a scale other than the optimal scale is provided by the asymptotic relative efficiency of the two tests. Let \mathcal{G} designate a family of scales, for each of which there is a different optimal test. Assume that the alternative hypothesis of a common association parameter $\theta_{sk} = \theta_s$ over all strata is true for scale $g_s(\pi) \in \mathcal{G}$, so that $X_{A(g_s)}^2$ is the asymptotically locally most powerful test. To simplify notation, designate $X_{A(g_s)}^2$ as $X_{A(s)}^2$. Then, let $g_r(\pi)$ be any other scale, $g_r(\pi) \in \mathcal{G}$, so that $\theta_{rj} \neq \theta_{rk}$ for some pair of strata $1 \leq j < k \leq K$. The two competing statistics for a test of association are

$$X_{A(\ell)}^2 = \frac{T_\ell^2}{\widehat{V}(T_\ell)}, \quad \ell = r, s, \quad (4.100)$$

where

$$T_\ell = \sum_k \widehat{w}_{\ell k} (p_{1k} - p_{2k}), \quad \ell = r, s, \quad (4.101)$$

$$\widehat{V}(T_\ell) = \widehat{V}(T_\ell | H_0) = \sum_k \widehat{w}_{\ell k}^2 \widehat{V}_{0k}, \quad \ell = r, s, \quad (4.102)$$

and

$$\widehat{w}_{\ell k} = \frac{1}{g'_\ell(p_k) \widehat{V}_{0k}}, \quad \ell = r, s. \quad (4.103)$$

Let $ARE(X_{A(r)}^2, X_{A(s)}^2)$ designate the ARE of $X_{A(r)}^2$ to $X_{A(s)}^2$ when the alternative hypothesis H_{1A} is true for scale $g_s(\pi)$ so that there is a common parameter $\theta_s \neq \theta_0$ for all strata. Since the tests are each based on weighted sums of the differences in proportions, T_r and T_s are each asymptotically normally distributed so that $ARE(X_{A(r)}^2, X_{A(s)}^2) = ARE(T_r, T_s)$. From (3.80) in Chapter 3, then

$$ARE(T_r, T_s) = \frac{Eff(T_r | \theta_s)}{Eff(T_s | \theta_s)} = \frac{\left[\left(\frac{dE(T_r | \theta)}{d\theta} \right)^2 \middle/ V(T_r) \right]}{\left[\left(\frac{dE(T_s | \theta)}{d\theta} \right)^2 \middle/ V(T_s) \right]} \Big|_{\theta=\theta_0}. \quad (4.104)$$

First consider the efficiency of the optimal test T_s . Substituting \hat{w}_{sk} into the expression for T_s , then

$$E(T_s | \theta_s) = E \left[\sum_k \frac{p_{1k} - p_{2k}}{g'_s(p_k) \hat{V}_{0k}} \right] = E \left[\sum_k \frac{g'_s(p_k) (p_{1k} - p_{2k})}{g'_s(p_k)^2 \hat{V}_{0k}} \right]. \quad (4.105)$$

From Slutsky's convergence theorem (A.45), since $g'_s(p_k) \xrightarrow{p} g'_s(\pi_k)$, then asymptotically

$$E[g'_s(p_k) (p_{1k} - p_{2k})] \cong g'_s(\pi_k) (\pi_{1k} - \pi_{2k}) = \theta_s \quad \forall k. \quad (4.106)$$

Likewise, since $\hat{V}_{0k} \xrightarrow{p} \sigma_{0k}^2$, then $g'_s(p_k)^2 \hat{V}_{0k} \xrightarrow{p} g'_s(\pi_k)^2 \sigma_{0k}^2$. Substituting each into the above yields

$$E(T_s | \theta_s) = \theta_s \sum_k \frac{1}{g'_s(\pi_k)^2 \sigma_{0k}^2}. \quad (4.107)$$

Thus,

$$\frac{dE(T_s | \theta_s)}{d\theta_s} = \sum_k \frac{1}{g'_s(\pi_k)^2 \sigma_{0k}^2} = \sum_k \frac{\sigma_{0k}^2}{[g'_s(\pi_k) \sigma_{0k}^2]^2} = \sum_k w_{sk}^2 \sigma_{0k}^2. \quad (4.108)$$

Again using Slutsky's convergence theorem, since $\hat{w}_{sk} \xrightarrow{p} w_{sk}$ and $\hat{V}_{0k} \xrightarrow{p} \sigma_{0k}^2$, then

$$\hat{V}(T_s | \theta_s) \xrightarrow{p} V(T_s | \theta_s) = \sum_k w_{sk}^2 \sigma_{0k}^2. \quad (4.109)$$

Therefore,

$$Eff(T_s | \theta_s) = \frac{\left[\sum_k w_{sk}^2 \sigma_{0k}^2 \right]^2}{\sum_k w_{sk}^2 \sigma_{0k}^2} = \sum_k w_{sk}^2 \sigma_{0k}^2. \quad (4.110)$$

Now consider the efficiency of the alternative test T_r . Since we assume that $\theta_{sk} = \theta_s \forall k$, this implies that θ_{rk} is not constant for all strata so that T_r will be suboptimal. Evaluating the efficiency of T_r under this assumption yields

$$E(T_r) = E \left[\sum_k \frac{p_{1k} - p_{2k}}{g'_r(\pi_k) \sigma_{0k}^2} \right]. \quad (4.111)$$

However, from (4.106) the alternative of a constant difference on scale $g_s(\pi)$ for all strata, in turn, implies that asymptotically

$$E(p_{1k} - p_{2k} | \theta_s) \cong \frac{\theta_s}{g'_s(\pi_k)}. \quad (4.112)$$

Therefore,

$$E(T_r | \theta_s) = \theta_s \left[\sum_k \frac{1}{g'_s(\pi_k) g'_r(\pi_k) \sigma_{0k}^2} \right], \quad (4.113)$$

and

$$\frac{dE(T_r | \theta_s)}{d\theta_s} = \left[\sum_k \frac{1}{g'_s(\pi_k) g'_r(\pi_k) \sigma_{0k}^2} \right] = \sum_k w_{rk} w_{sk} \sigma_{0k}^2. \quad (4.114)$$

Since

$$V(T_r | \theta_s) = \sum_k w_{rk}^2 \sigma_{0k}^2, \quad (4.115)$$

then

$$ARE(T_r, T_s | \theta_s) = \frac{[\sum_k w_{rk} w_{sk} \sigma_{0k}^2]^2}{[\sum_k w_{rk}^2 \sigma_{0k}^2][\sum_k w_{sk}^2 \sigma_{0k}^2]} = \rho_{rs}^2, \quad (4.116)$$

where

$$\rho_{rs} = \text{corr}(T_r, T_s). \quad (4.117)$$

These expressions involve the asymptotic null variance σ_{0k}^2 , that in turn involves the sample sizes within each stratum, the $\{n_{ik}\}$. However, we can factor N from the expression for σ_{0k}^2 by substituting $E(n_{ik}) = N\zeta_k \xi_{ik}$, where $\zeta_k = E(N_k/N)$ are the expected stratum sample fractions, and $\xi_{ik} = E(n_{ik}/N_k)$ are the expected group sample fractions within the k th stratum. Then

$$\sigma_{0k}^2 = \frac{\pi_k (1 - \pi_k)}{N \zeta_k} \left(\frac{1}{\xi_{1k}} + \frac{1}{\xi_{2k}} \right) = \frac{\phi_{0k}^2}{N}, \quad (4.118)$$

where ϕ_{0k}^2 does not involve the value of N . Thus, for each scale $\ell = r, s$, the weights $\{w_{\ell k}\}$ are proportional to

$$\tilde{w}_{\ell k} = \frac{1}{g'_\ell(\pi_k) \phi_{0k}^2}. \quad (4.119)$$

Substituting into (4.116), the resulting expression for $ARE(T_r, T_s)$ is a function only of the $\{\pi_k\}$, $\{\zeta_k\}$, and $\{\xi_{ik}\}$.

For a given set of data, the ARE can be consistently estimated as the sample correlation of the two test statistics based on the estimated weights and variances, expressed as

$$\widehat{ARE}(T_r, T_s | \theta_s) = \frac{[\sum_k \widehat{w}_{rk} \widehat{w}_{sk} \widehat{V}_{0k}]^2}{[\sum_k \widehat{w}_{rk}^2 \widehat{V}_{0k}][\sum_k \widehat{w}_{sk}^2 \widehat{V}_{0k}]} = \widehat{\rho}_{rs}^2. \quad (4.120)$$

The denominator is the product of the variances of the two test statistics, $\widehat{V}(T_r | H_0)$ and $\widehat{V}(T_s | H_0)$. The numerator can either be computed as the sum of cross products directly, or it can be simplified by noting that the product $(\widehat{w}_{rk} \widehat{w}_{sk} \widehat{V}_{0k})$ includes terms that cancel (see Problem 4.6.4).

Example 4.15 *Two Homogeneous Strata*

Again consider the data from Example 4.14 with two strata for which there is homogeneity of odds ratios and heterogeneity of the risk differences. If we assume that the limiting sample fractions are $\xi_{1k} = \xi_{2k} = 1/2$ for $k = 1, 2$, then

Stratum	π	ζ_k	ϕ_{0k}^2	Odds Ratio		Risk Difference	
				$g'_s(\pi)$	\tilde{w}_{sk}	$g'_r(\pi)$	\tilde{w}_{rk}
1	0.146	0.6	0.829	8.04	0.15	1	1.21
2	0.223	0.4	1.733	5.77	0.10	1	0.58

where $g'_s(\pi) = [\pi(1 - \pi)]^{-1}$ and $g'_r(\pi) = 1$. Therefore, asymptotically, regardless of the sample size,

$$\begin{aligned}
 & \text{ARE}(T_r, T_s | \theta_s) \\
 &= \frac{[(0.15 \times 1.21 \times 0.8289) + (0.1 \times 0.58 \times 1.733)]^2}{[(0.15)^2 (0.8289) + (0.1)^2 (1.733)] [(1.21)^2 (0.8289) + (0.58)^2 (1.733)]} \\
 &= 0.974.
 \end{aligned}$$

Therefore, with a large sample there is a loss of efficiency of 2.6% if we use the test designed to detect a common risk difference rather than the test designed to detect a common odds ratio that is optimal for this example.

Example 4.16 *Clinical Trial in Duodenal Ulcers (continued)*

For the ulcer clinical trial example, the terms entering into the computation of the covariance for each pair of tests $\{\hat{w}_{rk}\hat{w}_{sk}\hat{V}_{0k}\}$ are as follows:

Stratum	$\hat{w}_{rk}\hat{w}_{sk}\hat{V}_{0k}$		
	$RD, \log RR$	$RD, \log OR$	$\log OR, \log RR$
1	37.2453	22.1798	8.9716
2	13.5000	5.1429	3.1837
3	44.0000	22.4889	10.9946
<i>Total</i>	94.7453	49.8115	23.1498

Thus, the estimated AREs $\{\hat{\rho}_{rs}^2\}$ for each pair of tests are

<i>Scales</i> (r, s)	$\hat{V}(T_r)$	$\hat{V}(T_s)$	$\widehat{\text{Cov}}(T_r, T_s)$	$\widehat{\text{ARE}}(T_r, T_s)$
$RD, \log RR$	203.8863	44.9338	94.7453	0.9798
$RD, \log OR$	203.8863	12.1749	49.8115	0.9996
$\log OR, \log RR$	44.9338	12.1749	23.1498	0.9796

The tests for a common risk difference and for a common log odds ratio each have an ARE of about 0.98, indicating about a 2% loss of efficiency relative to the test for a common log relative risk. The ARE of the test for risk differences versus that for the log odds ratios is close to 1.0, neither test being clearly preferable to the other.

Example 4.17 *Religion and Mortality (continued)*

Likewise, for the association between religion and mortality, the estimated *AREs* for each pair of tests are

<i>Scales</i> (r, s)	$\widehat{ARE}(T_r, T_s)$
Risk difference, log Relative risk	0.4517
Risk difference, log Odds ratio	0.5372
log Odds ratio, log Relative risk	0.9825

Thus, there is a substantial loss in power using the risk differences relative to a test using either the relative risks or the odds ratios.

4.9 MAXIMIN-EFFICIENT ROBUST TESTS

Since the efficiency of the a test of partial association depends on the extent of homogeneity, the most efficient or powerful test is not known a priori. Therefore, in practice it would be preferable to use a test that is robust to the choice of scale, meaning one that has good power irrespective of whichever scale is, in fact, optimal. One approach to developing such a robust test is to choose a maximin-efficient test with respect to the family of tests. Two such tests are the Gastwirth scale robust test of association and the Wei-Lachin test of stochastic ordering.

4.9.1 Maximin Efficiency

Again consider a family of tests \mathcal{G} each of which is optimal under a specific alternative. For the Radhakrishna family we defined the family of tests \mathcal{G} based on the difference on some scale $g(\pi)$ for $g \in \mathcal{G}$. Herein we have restricted consideration to the family \mathcal{G} , consisting of the logit, log, and identity functions corresponding to the log odds ratio, log relative risk, and risk difference scales, respectively. This family, however, could be extended to include other functions, such as the arcsine, probit and square root scales that are explored as problems.

Let the test T_s corresponding to the scale $g_s \in \mathcal{G}$ be the optimal test within the family for a given set of data. Then let T_r be any other test corresponding to a different scale $g_r \in \mathcal{G}$ within the family. The test T_r may, in fact, be optimal for another set of data, but it is suboptimal for the data at hand. Thus, $ARE(T_r, T_s) < 1$ for $r \neq s$. However, the optimal test T_s is unknown a priori and whenever one prespecifies a particular test, one risks choosing a suboptimal test T_r with a pursuant loss of efficiency (power). Instead of prespecifying a particular test, a *maximin-efficient robust test (MERT)* can be defined as one that suffers the least loss in efficiency (power) irrespective of whichever member of the family is optimal for any given set of data.

Let Z_m designate such a *MERT* for the family \mathcal{G} . Formally, Z_m is chosen from a family of possible test statistics $Z \in \mathcal{M}$ such that Z_m maximizes the minimum *ARE* with respect to whichever member of the family \mathcal{G} is optimal. Then Z_m satisfies the

relationship

$$\sup_{Z \in \mathcal{M}} \inf_{g \in \mathcal{G}} \text{ARE}(Z, T_g) = \inf_{g \in \mathcal{G}} \text{ARE}(Z_m, T_g). \quad (4.121)$$

The expression on the right hand side is the minimum *ARE* of the *MERT* with respect to any member of the family \mathcal{G} . Thus, Z_m maximizes the minimum relative efficiency, regardless of which scale $g \in \mathcal{G}$ provides the test T_g that is optimal. Since T_s is the optimal test for the data at hand, then the *ARE* of the *MERT* is

$$\inf_{g \in \mathcal{G}} \text{ARE}(Z_m, T_g) = \text{ARE}(Z_m, T_s). \quad (4.122)$$

A comparable interpretation of maximin efficiency is in terms of power. The *MERT* Z_m suffers the least possible loss of power relative to the optimal test within the family \mathcal{G} . Thus, the test with maximin efficiency with respect to the family of alternatives is the test with minimax loss in power, that is, the test that minimizes the maximum loss in power compared to whichever test is optimal.

4.9.2 Gastwirth Scale Robust Test

Now let \mathcal{G} refer to the Radhakrishna family of tests, each of which is asymptotically most powerful for a restricted alternative hypothesis of the form $H_{1A(g)}$ in (4.79) for a specific scale $g(\pi) \in \mathcal{G}$. For a given set of data, let Z_g refer to the normal deviate test corresponding to the root of $X_{A(g)}^2$ in (4.97) for scale $g(\pi) \in \mathcal{G}$. Let (Z_r, Z_s) be the *extreme pair* of tests within the family, defined as

$$\rho_{rs} = \min_{r,s} (\rho_{i,k}) \quad (g_i, g_k) \in \mathcal{G}, \quad (4.123)$$

where $\rho_{rs}^2 = \text{ARE}(Z_r, Z_s) = \text{ARE}(T_r, T_s)$ from (4.116) is the *ARE* of Z_r to Z_s . Usually, $Z_r = \min_{g \in \mathcal{G}} (Z_g)$ and $Z_s = \sup_{g \in \mathcal{G}} (Z_g)$, so that Z_r is the test for the scale with the greatest heterogeneity among the strata, while Z_s is that for the scale with the greatest homogeneity among the strata with respect to the corresponding measures of association. Gastwirth (1985) then showed that if

$$\rho_{rg} + \rho_{sg} \geq 1 + \rho_{rs} \quad \forall (g \in \mathcal{G}), \quad g \neq r, \quad g \neq s, \quad (4.124)$$

then the maximin-efficient scale robust test (*MERT*) is obtained as a convex combination of the extreme pair of tests

$$Z_m = \frac{Z_r + Z_s}{[2(1 + \rho_{rs})]^{1/2}} \quad (4.125)$$

that is asymptotically distributed as standard normal under H_0 . The maximin efficiency of the *MERT* then is

$$\inf_{g \in \mathcal{G}} \text{ARE}(Z_m, Z_g) = \frac{1 + \rho_{rs}}{2}, \quad (4.126)$$

meaning that the *ARE* of the *MERT* Z_m relative to the unknown optimal test within the family is at least this quantity.

Gastwirth (1966) also shows that if the condition (4.124) does not hold, then the *MERT* still exists but it must be obtained as a linear combination of the Z_g ,

$$Z_m = \sum_{g \in \mathcal{G}} a_g Z_g \quad (4.127)$$

with coefficients $\{a_g\}$ that satisfy a set of constraints. For a family of three tests as herein, these constraints are

$$\begin{aligned} a_1(1 - \rho_{12}) - a_2(1 - \rho_{12}) + a_3(1 - \rho_{23}) &= 0 \\ a_1(1 - \rho_{13}) + a_2(\rho_{12} - \rho_{23}) - a_3(1 - \rho_{23}) &= 0 \end{aligned} \quad (4.128)$$

and

$$\sum_{i=1}^3 \sum_{k=1}^3 a_i a_k \rho_{ik} = 1. \quad (4.129)$$

In this case, the coefficients must be solved by an iterative procedure.

Example 4.18 *Clinical Trial in Duodenal Ulcers (continued)*

From Example 4.16, the extreme pair of scales are the log relative risk (R) and log odds ratio (O) with correlation $\hat{\rho}_{R,O} = 0.98976$. The other pairs of correlations are $\hat{\rho}_{R,D} = 0.98987$ and $\hat{\rho}_{O,D} = 0.99978$. However,

$$(\hat{\rho}_{R,D} + \hat{\rho}_{O,D}) = 1.98965 < (1 + \hat{\rho}_{R,O}) = 1.98976$$

and the condition in (4.124) is not satisfied. Therefore, the *MERT* cannot be readily computed using the convex combination in (4.125). Rather, the iterative computation in (4.128) would be required.

Example 4.19 *Religion and Mortality (continued)*

From Example 4.17, the extreme pair of scales are the relative risk and risk difference with $\hat{\rho}_{R,D} = 0.67205$. The other pairs of correlations are $\hat{\rho}_{R,O} = 0.99123$ and $\hat{\rho}_{O,D} = 0.73297$ and the condition in (4.124) is satisfied:

$$(\hat{\rho}_{R,O} + \hat{\rho}_{O,D}) = 1.7242 > (1 + \hat{\rho}_{R,D}) = 1.67205.$$

Thus, the *MERT* can be readily computed as

$$Z_m = \frac{\sqrt{4.11535} + \sqrt{1.23176}}{\sqrt{2(1.67205)}} = 1.7162$$

with two-sided $p \leq 0.087$.

4.9.3 Wei-Lachin Test of Stochastic Ordering

The efficiency of the test of association is highly dependent on the chosen scale because the alternative hypothesis $H_{1A(g)}$ in (4.79) specifies that there is a common difference within each stratum on that scale, as depicted in Figure 4.4A. Thus, the

test is directed to a highly restricted subset of the general K -dimensional omnibus parameter space, and many meaningful values are excluded, such as where there is a risk difference of 0.2 in one stratum and 0.4 in another. In such cases there may be some quantitative differences in the measures of association among strata but no qualitative differences, meaning that the *direction* of the association is consistent among strata.

Therefore, another way to derive a more robust test is to specify a less restrictive alternative hypothesis. In an entirely different setting, that of a multivariate rank test for repeated measures, Wei and Lachin (1984) and Lachin (1992a) suggested a test of stochastic ordering that is directed toward the alternative hypothesis of a common *qualitative* degree of association on some scale rather than a strictly common quantitative value on that scale. First consider a generalized one-sided upper-tail test. In this case, the alternative hypothesis of *stochastic ordering* is

$$H_{1S}: \pi_{1k} \geq \pi_{2k} \quad (k = 1, \dots, K) \quad \text{and} \quad \pi_{1k} > \pi_{2k} \text{ for some } 1 \leq k \leq K. \quad (4.130)$$

This alternative specifies that the probabilities in group 1 are at least as great as those in group 2 for all strata, and are strictly greater for some. For a measure of association for some scale $g(\pi) \in \mathcal{G}$, as employed above for the Radhakrishna family, then using a simplified notation, this specifies that

$$H_{1S} \Leftrightarrow \theta_k \geq 0 \text{ for } \forall k \quad (4.131)$$

with a strict inequality for at least one k . Thus, it is sufficient to employ a test based on the risk differences.

This test can also be used to conduct a two-sided test of stochastic ordering for which the alternative hypothesis is stated as

$$H_{1S}: \pi_{ik} \geq \pi_{(3-i)k}, \quad i = 1 \text{ or } 2; k = 1, \dots, K, \quad (4.132)$$

with a strict inequality for at least one k . This is a two-sided specification that the probabilities in one group (either $i = 1$ or 2) are larger than those in the other group. Equivalently, the alternative specifies that

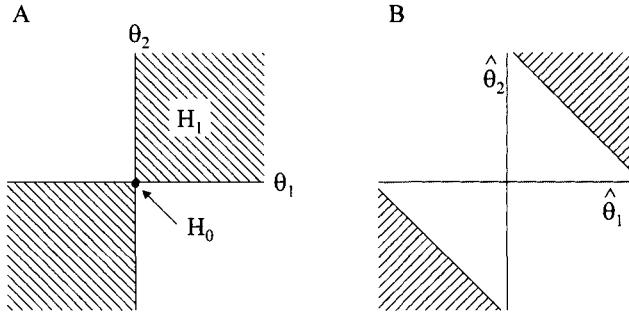
$$H_{1S} \Leftrightarrow (\theta_k \geq 0 \text{ for } \forall k) \text{ or } (\theta_k \leq 0 \text{ for } \forall k) \quad (4.133)$$

with a strict inequality for at least one k .

For the case of two strata or measures, Figure 4.5A shows the null and alternative hypothesis spaces in terms of the risk differences $\{RD_k\}$ in the two strata. The test of stochastic ordering is directed toward all points in quadrants I and III for which the two differences are either both positive or both negative. If either difference is zero, but not both, the convention used is to include that point in the alternative hypothesis parameter space.

Various authors have considered tests for this problem. For the case of multivariate normal variates with estimated variances and covariances, Perlman (1969) derives the computationally tedious likelihood ratio test of stochastic ordering; see also Kudo

Fig. 4.5 The null (H_0) and alternative (H_1) hypotheses for the test of stochastic ordering with two parameters $\theta = (\theta_1, \theta_2)$ (A) and the corresponding rejection region (B).



(1963). Tang et al. (1989) describe a less tedious approximation to Perlman's test that follows a simplified chi-bar-squared distribution.

In the setting of a stratified analysis of 2×2 tables, the Wei and Lachin (1984) test is based on the simple unweighted sum (or the unweighted mean) of the risk differences among strata,

$$X_S^2 = \frac{\sum_k (p_{1k} - p_{2k})]^2}{\sum_k \widehat{V}_{0k}}, \quad (4.134)$$

which is asymptotically distributed as χ_1^2 under H_0 . For example, consider the one-sided alternative hypothesis that the risk differences for two strata fall in the positive orthant, or the subhypothesis in H_{1S} that $RD_k \geq 0$ for $k = 1, 2$. This alternative hypothesis can be viewed as consisting of the family \mathcal{H} of all possible projections in the positive orthant. For each projection $h \in \mathcal{H}$ there is a corresponding optional test statistic of the form

$$T_h = w_{h_1} \widehat{RD}_1 + w_{h_2} \widehat{RD}_2. \quad (4.135)$$

The Wei-Lachin test, however, is based on

$$T_S = \widehat{RD}_1 + \widehat{RD}_2 \quad (4.136)$$

using unit (equal) weights.

Frick (1995) showed that the Wei-Lachin test is a maximin-efficient test in the following sense. For a vector of statistics $\widehat{\theta}$ with covariance matrix Σ as in Section

4.7, if the vector $\mathbf{J}'\Sigma$ is positive, \mathbf{J} being the unit vector, then the Wei-Lachin test is maximin-efficient with respect to the family \mathcal{H} of projections in the positive or negative orthant. For independent 2×2 tables, $\widehat{\boldsymbol{\theta}} = (\widehat{RD}_1 \ \widehat{RD}_2)^T$ with $\Sigma = \text{diag}(\sigma_{01}^2, \sigma_{02}^2)$, this condition obviously applies. Therefore, the Wei-Lachin test on average minimizes the loss in power relative to the optimal test corresponding to the true alternative hypothesis point (RD_1, RD_2) that lies on a particular projection in the positive or negative orthant.

For the bivariate example employed in Figure 4.2B for the omnibus test, and in Figure 4.4B for the test of association, Figure 4.5B presents the rejection region for the test of stochastic ordering. Because the test is based on a linear combination of the two statistics, the rejection region is defined by the simple sum of sample statistics satisfying $(\widehat{\theta}_1 + \widehat{\theta}_2)^2 / V(\widehat{\theta}_1 + \widehat{\theta}_2) = 3.841 = \chi^2_{1(0.95)}$ for $\alpha = 0.05$. This yields a line of rejection at a 135° angle to the X -axis. This differs from the line of rejection for the test of association that is tilted toward the origin for whichever $\widehat{\theta}_k$ has smaller variance, that is, whichever is more precise.

Example 4.20 *Clinical Trial in Duodenal Ulcers (continued)*

For the ulcer clinical trial the sum of the risk differences is $\sum_k \widehat{RD}_k = (-0.04458 + 0.30556 + 0.24506) = 0.50604$ and the sum of the variances estimated under the null hypothesis is $\sum_k \widehat{V}_{0k} = (0.01086 + 0.04586 + 0.01111) = 0.06783$. Thus, the Wei-Lachin test is

$$X_S^2 = \frac{(0.50604)^2}{0.06783} = 3.78$$

with $p \leq 0.052$. In comparison, the test of association directed toward a common risk difference yields $X_A^2 = 2.97$ with $p \leq 0.085$.

Example 4.21 *Religion and Mortality (continued)*

For the observational study of religion and mortality, $\sum_k \widehat{RD}_k = -0.33384$, $\sum_k \widehat{V}_{0k} = 0.02812$, and $X_S^2 = 3.96$ with $p \leq 0.0465$. In comparison, the test of association for a common risk difference yields $p \leq 0.27$ and the Gastwirth scale robust *MERT* yields $p \leq 0.086$.

4.9.4 Comparison of Weighted Tests

The omnibus test, all the members of the Radhakrishna family of weighted Cochran-Mantel-Haenszel tests of no partial association, the Gastwirth *MERT*, and the Wei-Lachin test of stochastic ordering are all tests of the global null hypothesis for the K strata in (4.48). The tests differ with respect to the alternative hypotheses H_{1O} , H_{1A} , and H_{1S} as depicted in Figures 4.2A, 4.4A and 4.5A, respectively. Each test will have greater power than the others in specific instances. When the $\{\theta_k\}$ are homogeneous on the prespecified scale, or nearly so, the test of association will be more powerful than either the omnibus test or the test of stochastic ordering, but not necessarily otherwise. Similarly, when the $\{\theta_k\}$ are homogeneous for one scale within a family of tests $g \in \mathcal{G}$, or nearly so, then the Gastwirth scale robust *MERT*

will be more powerful. Conversely, when the $\{\theta_k\}$ fall along a projection in the positive or negative orthants (quadrants I and III), the test of stochastic ordering will tend to be more powerful, especially if the corresponding projection under H_{1S} is not close to the projection of equality under H_{1A} for which the test of association is optimal. Finally, when the $\{\theta_k\}$ fall in some other orthant where some of the $\{\theta_k\}$ differ in sign (quadrants II and IV), then the omnibus test will tend to be more powerful than the others. See Lachin (1992a, 1996) for further comparison of these tests in the analysis of repeated measures.

In practice, there is a trade-off between the power robustness of the omnibus test to detect group differences under the broadest possible range of alternatives, versus the increased efficiency of the other tests to detect systematic differences between the groups under specific alternatives. As one compares the omnibus test, the test of stochastic ordering, and the test of association, in turn, there is decreasing robustness to a range of alternative hypotheses, but increasing power to detect specific restricted alternatives. In general, the Wei-Lachin test will have good power for alternatives approaching a constant difference on some scale. It will also have good power for alternatives where there is some heterogeneity but the risk differences are all in the same direction. In fact, X_S^2 is not nearly as sensitive to heterogeneity as is the test of association.

4.10 RANDOM EFFECTS MODEL

All of the methods described previously in this chapter are based on a fixed effects model. This model explicitly specifies that $E(\hat{\theta}_k) = \theta$ for $\forall k$ or that there is a common measure of association on some scale under the alternative hypothesis. This model specification is equivalent to the null hypothesis of homogeneity H_{0H} specified in (4.57) that is tested by the test of homogeneity on $K-1$ *df*. If heterogeneity is observed, it could arise for either of two reasons.

The first is that the fixed effects model has been misspecified in some respect. Perhaps there is homogeneity on some scale other than that specified for the analysis. Alternatively, perhaps an additional covariate must be adjusted for to yield homogeneity. For example, if we first adjust for ethnicity and find heterogeneity of odds ratios, then perhaps an adjustment for gender and ethnicity simultaneously would yield homogeneity over strata.

The second possibility is that the fixed effects model simply does not hold, meaning that there is some *extra-variation* or *overdispersion* due to random differences among strata. This extra-variation leads to the formulation of a random effects model.

4.10.1 Measurement Error Model

The simplest random effects model is a simple measurement error model, where a quantitative variable such as the level of serum cholesterol is measured with random

error and where the true cholesterol value varies at random from one subject to the next. These assumptions can be expressed in a *two-stage model* as follows.

Consider that we have a sample of independent observations $\{y_i\}$, $i = 1, \dots, N$. At the first stage of the model we assume that $y_i = v_i + \varepsilon_i$, where $E(y_i) = v_i$ is the true value that varies at random from one subject to the next. The conditional distribution $f(y_i|v_i)$ is determined by the value of v_i and the form of the distribution of the errors. For example, if $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, then $f(y_i|v_i)$ is $N(v_i, \sigma_\varepsilon^2)$. The second or random stage of the model then specifies that the v_i are randomly distributed in the population with some *mixing* distribution $v_i \sim f(v)$, where $E(v_i) = \mu$ and $V(v_i) = \sigma_v^2$. Thus, unconditionally the $\{y_i\}$ are distributed as $f(y) = \int_v f(y|v)f(v) dv$. Here we use f to denote the distribution of any random variable so that $f(y)$, $f(y|v)$, and $f(v)$ need not be the same distribution.

As in the usual fixed effects model, we also assume that the random errors are distributed as $\varepsilon_i \sim f(\varepsilon)$ with some distribution f , where $E(\varepsilon_i) = 0$, and $V(\varepsilon_i) = \sigma_\varepsilon^2$ for $\forall i$. In addition, we assume that $\varepsilon_i \perp v_i$ (independent of) $\forall i$ meaning that the random errors are statistically independent of the random conditional expectations. Note that the corresponding simple fixed effects model with only an overall mean (no strata or covariate effects) simply specifies that all observations share the same expectation such that $y_i = \mu + \varepsilon_i$ with $\varepsilon_i \sim h(\varepsilon)$, $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma_\varepsilon^2$ for all observations.

Using this random effects model specification, we then wish to estimate the moments of the mixing distribution μ and σ_v^2 and the variance of the errors σ_ε^2 . Intuitively, since the conditional expectations $\{v_i\}$ are assumed to be independent of the random errors $\{\varepsilon_i\}$, then $\sigma_y^2 = \sigma_v^2 + \sigma_\varepsilon^2$, or the total variation among the observed $\{y_i\}$ can be partitioned into the variation among the true values plus the variation among the random errors. This can be formally shown from the well-known result in (A.6), which in this measurement error model yields

$$\begin{aligned} V(Y) &= \sigma_y^2 = E(y - \mu)^2 = E[V(y|v)] + V[E(y|v)] \\ &= E\left[E(y - v)^2 | v\right] + V[v] \\ &= E\left[E(\varepsilon^2 | v)\right] + V[v]. \end{aligned} \quad (4.137)$$

If the errors have constant variance for all observations such that $E\left[E(\varepsilon^2 | v)\right] = \sigma_\varepsilon^2$ independent of v , then

$$V(Y) = \sigma_\varepsilon^2 + \sigma_v^2. \quad (4.138)$$

This also demonstrates the principle of partitioning of variation as in (A.5) since $\sigma_y^2 = \sigma_\varepsilon^2 + \sigma_v^2$ specifies that

$$E(y - \mu)^2 = E(y - v)^2 + E(v - \mu)^2. \quad (4.139)$$

The mean μ is readily estimated from \bar{y} , and the variance of Y from the usual sample estimate s_y^2 on $N-1$ df. If we can obtain an estimate of one of the variance components, usually $\hat{\sigma}_\varepsilon^2$, then we can obtain the other, usually $\hat{\sigma}_v^2$, by subtraction. For quantitative measurements, such as a laboratory assay for serum cholesterol,

these variance components are readily estimated from a set of independent duplicate measurements using moment estimators obtained from the expected mean squares of an analysis of variance (cf. Fleiss, 1986), or using restricted maximum likelihood or other methods (see Harville, 1977).

Such two-stage random effects models can be viewed as an application of what is often called the *NIH model*, a device pioneered by the early group of biostatisticians at the National Institutes of Health (Cornfield, Mantel, Haenszel, and Greenhouse, among others) that was never explicitly published. The NIH model was originally employed for an inference about the mean slope over time in a sample of subjects, each of whom has a unique slope (v) that is estimated from a set of repeated measurements over time. Within the population of subjects, these slopes follow a distribution with overall mean μ and variance σ_v^2 . The key is to then employ a moment estimator for the mixing distribution variance component, which is then used to obtain the total variance of the estimated mean slope.

4.10.2 Stratified-Adjusted Estimates from Multiple 2×2 Tables

DerSimonian and Laird (1986) applied these ideas to develop a random effects model for the analysis of multiple 2×2 tables. Their objective was to obtain an overall stratified-adjusted assessment of the treatment effect from a *meta-analysis* of many studies where there is some heterogeneity, or extra-variation or overdispersion, among studies.

Under a random effects model we now assume that the true measure of association for some scale $\theta_k = g(\pi_{1k}) - g(\pi_{2k})$ varies from stratum to stratum. Thus, at the first stage of the model we assume that

$$\hat{\theta}_k = \theta_k + \varepsilon_k, \quad (4.140)$$

where we again assume that

$$E(\varepsilon_k) = 0, \quad V(\varepsilon_k) = \sigma_{\varepsilon_k}^2 \quad \text{for } k = 1, \dots, K. \quad (4.141)$$

Note that $V(\varepsilon_k) = E(\hat{\theta}_k - \theta_k)^2 = \sigma_{\hat{\theta}_k}^2$, and thus the variance of the estimate $\sigma_{\hat{\theta}_k}^2$ is equivalent to the variance of the random errors $\sigma_{\varepsilon_k}^2$. Unlike the simple measurement error model, here the variance of the estimate (of the errors) is assumed to vary from unit to unit, or from stratum to stratum.

At the second stage we assume some mixing distribution

$$\theta_k \sim f(\theta; \mu_\theta, \sigma_\theta^2) \quad (4.142)$$

with

$$E(\theta_k) = \mu_\theta, \quad V(\theta_k) = \sigma_\theta^2 \quad (4.143)$$

and where $\varepsilon_k \perp \theta_k$. These variance components can be expressed as

$$\sigma_\theta^2 = E(\theta_k - \mu_\theta)^2 = V[E(\hat{\theta}_k | \theta_k)] \quad (4.144)$$

$$\sigma_{\hat{\theta}_k}^2 = E(\hat{\theta}_k - \theta_k)^2 = E[V(\hat{\theta}_k | \theta_k)],$$

where $\sigma_\theta^2 \equiv \sigma_v^2$ and $\sigma_{\hat{\theta}_k}^2 \equiv \sigma_\epsilon^2$ in the measurement error example of Section 4.10.1. Therefore, *unconditionally*,

$$V(\hat{\theta}_k) = \sigma_\theta^2 + \sigma_{\hat{\theta}_k}^2. \quad (4.145)$$

If the variance component $\sigma_\theta^2 = 0$, then this implies that the fixed effects model is appropriate. On the other hand, if $\sigma_\theta^2 > 0$, then there is some overdispersion relative to the fixed effects model. In this case, a fixed effects analysis yields stratified-adjusted tests and estimates for which the variance is underestimated because the model assumes that $\sigma_\theta^2 = 0$.

A test of homogeneity in effect provides a test of the null hypothesis H_{0H} : $\sigma_\theta^2 = 0$ versus the alternative H_{1H} : $\sigma_\theta^2 > 0$. If this test is significant, then a proper analysis using the two-stage random effects model requires that we estimate the between-stratum variance component σ_θ^2 . This is readily done using a simple moment estimator derived from the test of homogeneity.

Cochran's test of homogeneity $X_{H,C}^2$ in (4.67) can be expressed as a weighted sum of squares $\sum_k \tau_k (\hat{\theta}_k - \hat{\mu}_\theta)^2$, where $\hat{\mu}_\theta$ is the *MVLE* of the mean measure of association obtained under the fixed effects model and $\hat{\tau}_k$ is the inverse of the estimated variance of the estimate. Clearly, the expected value $E(X_{H,C}^2)$ under the alternative hypothesis H_{1H} will be some function of the variance between strata, σ_θ^2 . To obtain this expectation, we first apply the principle of partitioning of sums of squares described in Section A.1.3.

Treating the $\{\tau_k\}$ as fixed (known), then from (A.4), the sum of squares of each estimate about the overall mean can be partitioned about the estimated mean as

$$\sum_k \tau_k (\hat{\theta}_k - \mu_\theta)^2 = \sum_k \tau_k (\hat{\theta}_k - \hat{\mu}_\theta)^2 + \sum_k \tau_k (\hat{\mu}_\theta - \mu_\theta)^2, \quad (4.146)$$

so that

$$X_{H,C}^2 = \sum_k \tau_k (\hat{\theta}_k - \hat{\mu}_\theta)^2 = \sum_k \tau_k (\hat{\theta}_k - \mu_\theta)^2 - \sum_k \tau_k (\hat{\mu}_\theta - \mu_\theta)^2. \quad (4.147)$$

Since $E(\hat{\theta}_k - \mu_\theta)^2 = V(\hat{\theta}_k)$ in (4.145), then the expected value of the test statistic is

$$E(X_{H,C}^2) = \sum_k \tau_k V(\hat{\theta}_k) - V(\hat{\mu}_\theta) \left(\sum_k \tau_k \right). \quad (4.148)$$

In practice, we would use the estimated weights $\{\hat{\tau}_k\}$, as in Section 4.6.2. However, since $\hat{\tau}_k \xrightarrow{P} \tau_k$, then from Slutsky's theorem the above still applies.

Using the unconditional variance of each $\hat{\theta}_k$ in (4.145), the first term on the right hand side of (4.148) is

$$\sum_k \tau_k V(\hat{\theta}_k) = \sum_k \tau_k (\sigma_\theta^2 + \sigma_{\hat{\theta}_k}^2). \quad (4.149)$$

For the second term on the right hand side of (4.148), note that the *MVLE* is obtained as $\hat{\mu}_\theta = \sum_k \omega_k \hat{\theta}_k$ using the *MVLE* weights $\omega_k = \tau_k / \sum_\ell \tau_\ell$, where $\tau_k = \sigma_{\hat{\theta}_k}^{-2}$ in

(4.34) is assumed known (fixed). Again using the unconditional variance of each $\hat{\theta}_k$, given the *MVLE* weights,

$$V(\hat{\mu}_\theta) = \frac{\sum_k \tau_k^2 (\sigma_{\hat{\theta}_k}^2 + \sigma_\theta^2)}{(\sum_k \tau_k)^2}. \quad (4.150)$$

Thus,

$$V(\hat{\mu}_\theta) \sum_k \tau_k = \frac{\sum_k \tau_k^2 (\sigma_{\hat{\theta}_k}^2 + \sigma_\theta^2)}{\sum_k \tau_k}, \quad (4.151)$$

so that

$$E(X_{H,C}^2) = \sum_k \tau_k (\sigma_{\hat{\theta}_k}^2 + \sigma_\theta^2) - \frac{\sum_k \tau_k^2 (\sigma_{\hat{\theta}_k}^2 + \sigma_\theta^2)}{\sum_k \tau_k}. \quad (4.152)$$

Noting that $\tau_k = \sigma_{\hat{\theta}_k}^{-2}$, it is readily shown that

$$E(X_{H,C}^2) = (K-1) + \sigma_\theta^2 \left[\sum_k \tau_k - \frac{\sum_k \tau_k^2}{\sum_k \tau_k} \right]. \quad (4.153)$$

From Slutsky's theorem, a consistent estimate of this expectation is provided upon substituting the estimated *MVLE* weights or the $\{\hat{\tau}_k\}$. This yields a consistent moment estimate for σ_θ^2 of the form

$$\hat{\sigma}_\theta^2 = \max \left[0, \frac{X_{H,C}^2 - (K-1)}{\sum_k \hat{\tau}_k - \frac{\sum_k \hat{\tau}_k^2}{\sum_k \hat{\tau}_k}} \right], \quad (4.154)$$

where the estimate is set to zero when the solution is a negative value.

Given the estimate of the between strata variance component $\hat{\sigma}_\theta^2$, we can then update the estimate of the mean $\hat{\mu}_\theta$ by a reweighted estimate using the unconditional variance of the estimate within each stratum:

$$\hat{V}(\hat{\theta}_k) = \hat{\sigma}_{\hat{\theta}_k}^2 + \hat{\sigma}_\theta^2. \quad (4.155)$$

The initial *MVLE* estimate is called the *initial estimate* and the reweighted estimate is called the *first-step iterative estimate*. The first-step revised weights are

$$\hat{\omega}_k^{(1)} = \frac{\hat{\tau}_k^{(1)}}{\sum_\ell \hat{\tau}_\ell^{(1)}} = \frac{\hat{V}(\hat{\theta}_k)^{-1}}{\sum_\ell \hat{V}(\hat{\theta}_\ell)^{-1}} = \frac{\left(\hat{\sigma}_{\hat{\theta}_k}^2 + \hat{\sigma}_\theta^2 \right)^{-1}}{\sum_\ell \left(\hat{\sigma}_{\hat{\theta}_\ell}^2 + \hat{\sigma}_\theta^2 \right)^{-1}}. \quad (4.156)$$

The reweighted estimate of the mean over all strata then is

$$\hat{\mu}_\theta^{(1)} = \sum_k \hat{\omega}_k^{(1)} \hat{\theta}_k, \quad (4.157)$$

with estimated variance

$$\widehat{V} \left(\widehat{\mu}_\theta^{(1)} \right) = \sum_k \left(\widehat{\omega}_k^{(1)} \right)^2 \left(\widehat{\sigma}_{\theta_k}^2 + \widehat{\sigma}_\theta^2 \right). \quad (4.158)$$

From these quantities one can obtain a random effects confidence interval for the mean measure of association in the population.

DerSimonian and Laird also describe an iterative convergent solution using the EM-algorithm. Alternatively, the above process could be continued to obtain fully iterative estimates of the mean and its variance. The reweighted estimate of the mean $\widehat{\mu}_\theta^{(1)}$ would be used to recalculate the test of homogeneity, which, in turn, is used to update the estimate of the variance between strata $(\widehat{\sigma}_\theta^2)^{(2)}$. This updated estimate of the variance is used to obtain revised weights $\{\widehat{\omega}_k^{(2)}\}$ and then to obtain an updated estimate of the mean $\widehat{\mu}_\theta^{(2)}$, and so on. The iterative procedure continues until both the mean $\widehat{\mu}_\theta^{(m)}$ and its variance $\widehat{V} \left(\widehat{\mu}_\theta^{(m)} \right)$ converge to constants, say at the m th step. This approach is called the *fixed-point method* of solving a system of simultaneous equations. Alternatively, the two iterative estimates could be obtained simultaneously using other numerical procedures, such as Newton-Raphson (cf. Thisted, 1988).

For such calculations Lipsitz et al. (1994) have shown that the one-step estimates are often very close to the final iterative estimates, and that the mean square error of the one-step estimate also is close to that of the final iterative estimate. Thus, the first-step estimates are often used in practice.

This iteratively reweighted random effects estimate can also be described as an *empirical Bayes estimate* (see Robbins, 1964; Morris, 1983). The addition of the nonzero variance component between strata, $\widehat{\sigma}_\theta^2$, to the variance of the estimate to obtain the unconditional variance has the effect of adding a constant to all of the weights. Thus, the random effects (empirical Bayes) analysis "shrinks" the weights toward the average $1/K$, so that the resulting estimate is closer to the unweighted mean of the $\{\widehat{\theta}_k\}$ than is the *MVLE*. If the estimate of this variance component is zero, or nearly so, the random effects analysis differs trivially from the fixed effects analysis.

Thus, one strategy could be to always adopt a random effects model because there is usually some extra-variation or overdispersion, and if not, then $\widehat{\sigma}_\theta^2 \doteq 0$ and the fixed analysis will result. However, this could sacrifice power in those cases where a fixed effects model actually applies because even when $\sigma_\theta^2 = 0$, the estimate $\widehat{\sigma}_\theta^2$ will vary and a small value will still inflate the variance of the estimate of $\widehat{\mu}_\theta^{(1)}$. Thus, it is customary to first conduct a test of homogeneity and to conduct a random effects analysis only if significant heterogeneity is detected.

In theory, an asymptotically efficient test of the hypothesis $H_0: \mu_\theta = 0$ on a given scale under a random effects model could be obtained analogously to the Radhakrishna test of Section 4.7.2, that was derived under a fixed effects model. In practice, however, inference from a random effects analysis, such as in a meta-analysis, is usually based on the 95% confidence limits for the mean μ_θ .

Example 4.22 Clinical Trial in Duodenal Ulcers (continued)

For the ulcer drug trial example, the following is a summary of the computation of the one-step random effects estimate of the mean stratified-adjusted log odds ratio, where $\widehat{\theta}_k = \log \widehat{OR}_k$. As shown in Table 4.3, the *MVLE* under the fixed effects model is $\widehat{\theta} = 0.493$ with estimated variance $\widehat{V}(\widehat{\theta}) = 0.085$. The corresponding estimate of the assumed common odds ratio is $\widehat{OR} = 1.637$ with asymmetric 95% *C.I.* of (0.924, 2.903). The resulting test of homogeneity of the log odds ratios is $X^2_{H,C} = 4.58028$ with $p \leq 0.102$. Although not significant, the random effects analysis is presented for the purpose of illustration.

The moment estimate of the variance between strata is $\widehat{\sigma}_\theta^2 = 0.37861$. This is then used to obtain an updated (one-step) estimate of the mean μ_θ and its variance. The random effects analysis, contrasted to the original fixed effects (*MVLE*) analysis is as follows:

Stratum			<i>MVLE</i>		<i>Random Effects</i>		
	$\widehat{\theta}_k$	$\widehat{\sigma}_k^2$	$\widehat{\tau}_k$	$\widehat{\omega}_k$	$\widehat{V}(\widehat{\theta}_k)$	$\widehat{\tau}_k^{(1)}$	$\widehat{\omega}_k^{(1)}$
1	-0.185	0.188	5.319	0.454	0.567	1.765	0.409
2	1.322	0.894	1.118	0.095	1.273	0.786	0.182
3	1.001	0.189	5.277	0.451	0.568	1.760	0.408
<i>Total</i>			11.715	1.0		4.311	1.000

Through the addition of the variance component estimate $\widehat{\sigma}_\theta^2 = 0.37861$ to the unconditional variance of each estimate, the random effects weights for each stratum are now shrunk toward 1/3, such as from 0.454 to 0.409 for the first stratum. The resulting one-step reweighted estimate of the mean log odds ratio is $\widehat{\mu}_\theta^{(1)} = (-0.185 \times 0.409) + (1.322 \times 0.182) + (1.001 \times 0.408) = 0.574$ and $\widehat{\mu}_{OR}^{(1)} = 1.775$. The estimated variance of the log odds ratio is $\widehat{V}(\widehat{\mu}_\theta^{(1)}) = 1/4.311 = 0.2320$, which yields asymmetric 95% *C.I.* on μ_{OR} of (0.691, 4.563). The point estimate of the mean log odds ratio in the random effects model is slightly greater than the *MVLE* because the negative estimate in the first stratum is given less weight. However, the variance of the random effects estimate is slightly greater because of the allowance for the extra-variation between strata, so that the confidence limits are wider.

Table 4.9 presents a summary of the computations of the random effects estimates of the mean parameter μ_θ and its estimated large sample variance for the risk difference, relative risk, and odds ratio. Of the three scales, the relative risk estimates are the least affected by the random effects analysis. That is because the estimates of the relative risks among the strata showed the least heterogeneity. The corresponding estimate of the variance between strata in the log relative risks is $\widehat{\sigma}_\theta^2 = 0.06818$, which is smaller, relative to the variance of the estimates within strata, than the between-stratum variances for the other scales. Thus, the $\{\widehat{\omega}_k^{(1)}\}$ are minimally different in the fixed effects and random effects analyses of the relative risks.

The estimate of the variance between strata of the risk differences is $\widehat{\sigma}_\theta^2 = 0.02235$, which yields effects on the estimates similar to those observed for the log odds ratios.

Table 4.9 Random-effects model stratified-adjusted analysis of the ulcer clinical trial data from Example 4.1.

Measure, $\hat{\theta}_k$	Stratum			Mean	95% C.I.
	1	2	3	$\hat{\mu}_\theta^{(1)}$	
Risk difference	-0.045	0.306	0.245	0.142	-0.08, 0.37
$\hat{V}(\hat{\theta}_k)$	0.033	0.065	0.032	0.013	
$\hat{\omega}_k^{(1)}$	0.397	0.201	0.402		
log Relative risk	-0.111	0.523	0.515	0.284	
$\hat{V}(\hat{\theta}_k)$	0.136	0.235	0.122	0.050	
$\hat{\omega}_k^{(1)}$	0.372	0.215	0.413		
Relative risk				1.329	0.86, 2.06
log Odds ratio	-0.185	1.322	1.001	0.574	
$\hat{V}(\hat{\theta}_k)$	0.567	1.273	0.568	0.232	
$\hat{\omega}_k^{(1)}$	0.409	0.102	0.408		
Odds ratio				1.775	0.69, 4.56

Example 4.23 *Religion and Mortality (continued)*

For the observational study of religion and mortality, the estimates of the variances between strata are zero for the log odds ratios and the log relative risks, as reflected by the nonsignificant tests of homogeneity in Example 4.11. The estimate for the variance among strata for the risk differences is $\hat{\sigma}_\theta^2 = 0.00132$, which has a slight effect on the resulting estimates: the estimated mean being $\hat{\mu}_\theta^{(1)} = -0.040$ with $\hat{V}(\hat{\mu}_\theta^{(1)}) = 0.00117$ and 95% confidence limits of $(-0.107, 0.027)$, slightly wider than the fixed effects limits.

Example 4.24 *Meta-analysis of Effects of Diuretics on Pre-eclampsia*

Collins et al. (1985) present a meta-analysis of nine studies of the use of diuretics during pregnancy to prevent the development of pre-eclampsia. The data are presented in Table 4.10. Of the nine studies, three show an increase in the odds ratio of pre-eclampsia among those treated with diuretics, whereas the others show a decreased risk. The Cochran test of homogeneity of the log odds ratios yields $X^2 = 27.3$ on 8 *df* with $p \leq 0.0007$. The initial estimate of the variance in the log odds ratios between studies is $\hat{\sigma}_\theta^2 = 0.2297$. The one-step random effects estimate of the average log odds ratio is $\hat{\mu}_\theta^{(1)} = -0.517$ with estimated variance $\hat{V}(\hat{\mu}_\theta^{(1)}) = 0.0415$, which yields a point estimate of the average odds ratio of $\hat{\mu}_{OR}^{(1)} = 0.596$ with asymmetric 95% confidence limits $(0.400, 0.889)$ that is significant at the 0.05 level since the limits do not bracket 1.0.

Table 4.10 Meta-analysis of prevention of pre-eclampsia with diuretics during pregnancy. From Collins et al. (1985), reproduced with permission.

Study	Diuretics Group			Placebo Group			OR_k
	a_k	n_{1k}	p_{1k}	b_k	n_{2k}	p_{2k}	
1	14	131	0.107	14	136	0.103	1.043
2	21	385	0.055	17	134	0.127	0.397
3	14	57	0.246	24	48	0.500	0.326
4	6	38	0.158	18	40	0.450	0.229
5	12	1011	0.012	35	760	0.046	0.249
6	138	1370	0.101	175	1336	0.131	0.743
7	15	506	0.030	20	524	0.038	0.770
8	6	108	0.056	2	103	0.019	2.971
9	65	153	0.425	40	102	0.392	1.145

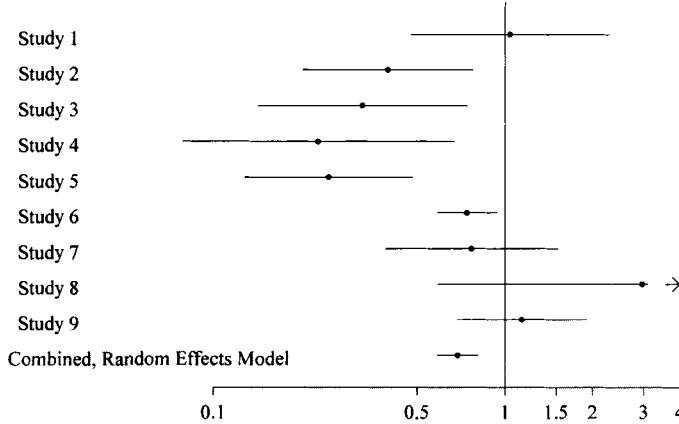
Applying the fixed-point iterative algorithm, wherein the estimate $\hat{\sigma}_\theta^2$ is used to obtain an updated estimate $\hat{\mu}_\theta^{(2)}$, and so on, requires 27 iterations to reach convergence. The final estimate of the mixing distribution variance is $\hat{\sigma}_\theta^2 = 0.1733$, somewhat less than the initial estimate. The resulting estimate of the mean log odds ratio is $\hat{\mu}_\theta = -0.513$ with estimated variance $\hat{V}(\hat{\mu}_\theta) = 0.0346$. The estimate of the mean is virtually unchanged, but that of the variance is slightly less. The final estimate of the mean odds ratio, therefore, is $\hat{\mu}_{OR} = 0.599$ with asymmetric 95% confidence limits (0.416, 0.862).

In a meta-analysis it is traditional to display the results as in Figure 4.6, which shows the odds ratio and 95% confidence limits within each study and in aggregate. For illustration, the figure displays the random effects model estimate of the aggregate combined odds ratio and its 95% confidence limits. Because the confidence interval does not bracket 1.0, the conclusion would be that the aggregate estimate is statistically significant at the 0.05 level.

4.11 POWER AND SAMPLE SIZE FOR TESTS OF ASSOCIATION

Wittes and Wallenstein (1987) describe the power function of the Mantel-Haenszel test for multiple 2×2 tables. Earlier, Gail (1973) presented the power function for a similar but different test, a Wald test based on the *MVLE*; and Birch (1964) described the conditional power function using the noncentrality parameter of the noncentral chi-square distribution. Because the Mantel-Haenszel test is asymptotically equivalent to Cochran's test, and since Cochran's test is a member of the Radhakrishna family, the Wittes-Wallenstein result can be derived more generally for the Radhakrishna family.

Fig. 4.6 Meta-analysis display of the odds ratio and 95% confidence limits on the log scale for studies of pre-eclampsia, and the random effects model combined analysis.



4.11.1 Power Function of the Radhakrishna Family

The Radhakrishna family provides an asymptotically efficient test for the assumed common measure of association θ , where $\theta_k = g(\pi_{1k}) - g(\pi_{2k}) = \theta \forall k$ under a fixed effects model. The test statistic T in (4.84) is obtained as a weighted linear combination of the risk differences within each stratum, the weights providing an asymptotically optimal test for a measure on the specified scale $g(\pi)$. The variance of the test statistic involves the variance of the risk differences σ_{0k}^2 , that involves the sample sizes within each stratum. As shown in (4.118), the variance can be factored as ϕ_{0k}^2/N using $E(n_{ik}) = N\zeta_k\xi_{ik}$, where $\zeta_k = E(N_k/N)$ are the expected stratum sample fractions and $\xi_{ik} = E(n_{ik}/N_k)$ the group sample fractions within the k th stratum. Thus,

$$\phi_{0k}^2 = \frac{\pi_k(1-\pi_k)}{\zeta_k\xi_{1k}\xi_{2k}}. \quad (4.159)$$

Therefore, the test statistic can be expressed as

$$X_{A(g)}^2 = \frac{T^2}{V(T|H_0)} = \frac{\sum_k [\tilde{w}_k (p_{1k} - p_{2k})]^2}{\sum_k \tilde{w}_k^2 \sigma_{0k}^2} = \frac{\tilde{T}^2}{V(\tilde{T}|H_0)} \quad (4.160)$$

using the weights

$$\tilde{w}_k = \frac{1}{\phi_{0k}^2 g'(\pi_k)}. \quad (4.161)$$

Under H_0 : $\pi_{1k} = \pi_{2k} = \pi_k$, then $\tilde{T} \stackrel{d}{\approx} \mathcal{N} \left(0, \sigma_{\tilde{T}_0}^2 \right)$, where $E \left(\tilde{T} \right) = 0$ and

$$\begin{aligned} V \left(\tilde{T} | H_0 \right) &= \sigma_{\tilde{T}_0}^2 = \sum_k \tilde{w}_k^2 \sigma_{0k}^2 = \sum_k \tilde{w}_k^2 \frac{\phi_{0k}^2}{N} \\ &= \frac{1}{N} \sum_k \frac{1}{\phi_{0k}^2 g'(\pi_k)^2} = \frac{\phi_{\tilde{T}_0}^2}{N}, \end{aligned} \quad (4.162)$$

where $\phi_{\tilde{T}_0}^2$ is the variance component remaining after factoring N from this expression for the variance under H_0 .

Under the omnibus alternative H_{1O} in (4.49) that some of the stratum-specific parameters differ from the null value, $\theta_k \neq \theta_0 = 0$ for some $1 \leq k \leq K$, then asymptotically $\tilde{T} \stackrel{d}{\approx} \mathcal{N} \left(\mu_{\tilde{T}}, \sigma_{\tilde{T}_1}^2 \right)$ with mean

$$\mu_{\tilde{T}} = \sum_k \tilde{w}_k (\pi_{1k} - \pi_{2k}) \cong \sum_k \frac{\theta_k}{\phi_{0k}^2 g'(\pi_k)^2}, \quad (4.163)$$

and variance

$$\sigma_{\tilde{T}_1}^2 = \sum_k \tilde{w}_k^2 \sigma_{1k}^2, \quad (4.164)$$

where $\sigma_{1k}^2 = V(p_{1k} - p_{2k} | H_1)$ is the variance of the risk difference under the alternative hypothesis within the k th stratum as in (2.26). This variance likewise can be factored as $\sigma_{1k}^2 = \phi_{1k}^2 / N$ with

$$\phi_{1k}^2 = \frac{1}{\zeta_k} \left(\frac{\pi_{1k} (1 - \pi_{1k})}{\xi_{1k}} + \frac{\pi_{2k} (1 - \pi_{2k})}{\xi_{2k}} \right). \quad (4.165)$$

Therefore, the asymptotic variance under the alternative is

$$\sigma_{\tilde{T}_1}^2 = \frac{1}{N} \sum_k \frac{\phi_{1k}^2}{[\phi_{0k}^2 g'(\pi_k)]^2} = \frac{\phi_{\tilde{T}_1}^2}{N}. \quad (4.166)$$

From the general equation for sample size and power (3.18), we then obtain

$$|\mu_{\tilde{T}}| = Z_{1-\alpha} \frac{\phi_{\tilde{T}_0}}{\sqrt{N}} + Z_{1-\beta} \frac{\phi_{\tilde{T}_1}}{\sqrt{N}} \quad (4.167)$$

and

$$\sqrt{N} |\mu_{\tilde{T}}| = Z_{1-\alpha} \left[\sum_k \frac{1}{\phi_{0k}^2 g'(\pi_k)^2} \right]^{1/2} + Z_{1-\beta} \left[\sum_k \frac{\phi_{1k}^2}{[\phi_{0k}^2 g'(\pi_k)]^2} \right]^{1/2}. \quad (4.168)$$

This expression can be solved for N to determine the sample size required to provide any specific level of power, and can be solved for $Z_{1-\beta}$ to determine the power of a test for any given sample size.

4.11.2 Power and Sample Size for Cochran's Test

Cochran's test is directed toward alternatives specified in terms of odds ratios using the logit link, $g(\pi) = \log[\pi(1 - \pi)]$, for which the optimal weights are given in (4.81). Factoring the total sample size N from this expression yields

$$w_k = \frac{N}{\phi_{0k}^2 g'(\pi_k)} = \frac{N}{\left(\frac{1}{\zeta_k \xi_{1k}} + \frac{1}{\zeta_k \xi_{2k}} \right)} \quad (4.169)$$

so that

$$\tilde{w}_k = \left(\frac{1}{\zeta_k \xi_{1k} \xi_{2k}} \right)^{-1} = \zeta_k \xi_{1k} \xi_{2k}. \quad (4.170)$$

Therefore, the variance component under the null hypothesis is

$$\phi_{\tilde{T}_0}^2 = \sum_k \tilde{w}_k^2 \phi_{0k}^2 = \sum_k \zeta_k \xi_{1k} \xi_{2k} \pi_k (1 - \pi_k). \quad (4.171)$$

Under the alternative, the expectation is

$$\mu_{\tilde{T}} = \sum_k \zeta_k \xi_{1k} \xi_{2k} (\pi_{1k} - \pi_{2k}), \quad (4.172)$$

with variance component

$$\phi_{\tilde{T}_1}^2 = \sum_k \tilde{w}_k^2 \phi_{1k}^2 = \sum_k (\zeta_k \xi_{1k} \xi_{2k}) [\xi_{2k} \pi_{1k} (1 - \pi_{1k}) + \xi_{1k} \pi_{2k} (1 - \pi_{2k})]. \quad (4.173)$$

Therefore, the expression for the power of the test is

$$Z_{1-\beta} = \frac{|\mu_{\tilde{T}}| \sqrt{N} - Z_{1-\alpha} \phi_{\tilde{T}_0}}{\phi_{\tilde{T}_1}}, \quad (4.174)$$

and the expression for the sample size required to provide power $1 - \beta$ is

$$N = \left[\frac{Z_{1-\alpha} \phi_{\tilde{T}_0} + Z_{1-\beta} \phi_{\tilde{T}_1}}{\mu_{\tilde{T}}} \right]^2. \quad (4.175)$$

To perform computations of the power or required sample size under a specific alternative hypothesis, one first specifies the expected sample fractions ζ_k , ξ_{1k} , and ξ_{2k} for each stratum ($1 \leq k \leq K$). For each stratum the expected control group probability π_{2k} is specified as well as the expected log odds ratio θ_k in that stratum. By inverting the expression for the odds ratio, the expected probability in the exposed or treated group is then obtained as

$$\pi_{1k} = \frac{\pi_{2k} e^{\theta_k}}{(1 - \pi_{2k}) + \pi_{2k} e^{\theta_k}}. \quad (4.176)$$

Although the test is asymptotically most powerful when there is absolute homogeneity of the odds ratios, $\theta_k = \theta$ for all strata, these expressions allow for the θ_k to differ systematically among strata under a fixed effects model without a common value for θ . Also, the variance components in (4.171) and (4.173) are obtained under this fixed effects model.

Example 4.25 *Three Strata with Heterogeneity*

Consider a nonrandomized study with three strata, allowing for some imbalances between groups among strata ($\xi_{1k} \neq \xi_{2k}$) and some heterogeneity of odds ratios as reflected by the following parameters under the alternative hypothesis:

Age Stratum	ζ_k	Stratum		Group Fractions		OR_k (e^{θ_k})	π_{1k}
		ξ_{1k}	ξ_{2k}	π_{2k}			
20–49	0.15	0.30	0.70	0.75	4.0	0.923	
50–69	0.50	0.60	0.40	0.70	3.2	0.882	
70–80	0.35	0.45	0.55	0.65	1.8	0.770	

As described above, the values $\{\pi_{1k}\}$ are obtained from the specification of the control group probabilities $\{\pi_{2k}\}$ and the odds ratio $\{OR_k\}$. For this example, $\mu_{\bar{T}} = 0.0377$, $\phi_{\bar{T}_0} = 0.2055$, and $\phi_{\bar{T}_1} = 0.2033$. For a test at $\alpha = 0.05$ (two-sided), in order to provide 90% power ($\beta = 0.10$), a total sample size $N = 306.8$ is required.

This calculation could be compared to the sample size required to provide 90% power in a marginal analysis. Weighting by the stratum-specific sample fractions, $\zeta_k \xi_{ik}$ yields overall group fractions $\xi_{1\bullet} = \sum_k \zeta_k \xi_{1k} = 0.5025$, $\xi_{2\bullet} = 0.4975$, and probabilities $\pi_{2\bullet} = \sum_k \zeta_k \xi_{2k} \pi_{2k} / \xi_{2\bullet} = 0.689$ and $\pi_{1\bullet} = \sum_k \zeta_k \xi_{1k} \pi_{1k} / \xi_{1\bullet} = 0.847$. The resulting marginal odds ratio is $OR_{\bullet} = 2.505$. Then using the equation for sample size for the test of two proportions in (3.37) yields $N = 294.4$.

Therefore, the marginal analysis appears to be more powerful and to require a smaller sample size than does the stratified analysis. This is caused by the heterogeneity of the odds ratios over strata and the treatment group imbalances over strata. As shown in Section 4.4.3, the marginal analysis is biased in this case. The stratified analysis, however, is unbiased, and the larger sample size is indeed required to provide an unbiased test with the desired level of power.

The impact of these factors is described in Table 4.11, which compares the sample sizes required for a stratified versus a marginal analysis where the parameters in this example are modified so as to have either homogeneity or heterogeneity coupled with either group by covariate imbalance or balance. Under homogeneity of odds ratios, the odds ratio within each stratum is assumed to be $OR_k = 2.52 \quad \forall k$, which is the weighted average odds ratio among the three strata. Also in the case of a balanced design, the sample fractions are assumed equal in each group ($\xi_{1k} = \xi_{2k}$). The computations in the first row of the table are those described above. Such computations demonstrate that a moderate degree of heterogeneity alone has little effect on the power of the stratified versus that of the marginal analysis. The important determinant of power is the average odds ratio. The introduction of a group by covariate imbalance appears to reduce the power of the stratified test due to the pursuant increase in the variance of the test because the term

$$\frac{1}{\xi_{1k}} + \frac{1}{\xi_{2k}} = \frac{1}{\xi_{1k}\xi_{2k}} \quad (4.177)$$

is a minimum for $\xi_{1k} = 0.5$.

Table 4.11 Required sample size for stratified and unstratified analysis with homogeneity or heterogeneity of odds ratios, and with balanced or unbalanced group fractions

<i>Odds Ratios Among Strata</i>	<i>Group Sample Fractions</i>	<i>Analysis Stratified</i>	<i>Analysis Marginal</i>
Heterogeneous:			
	Unbalanced	306.8	294.4
	Balanced	287.8	291.5
Homogeneous:			
	Unbalanced	305.6	287.3
	Balanced	292.4	293.6

Lachin and Bautista (1995) provide a detailed assessment of factors influencing the power of a stratified versus unstratified analysis. They considered three factors: (a) the degree of heterogeneity among the odds ratios $\{\theta_k\}$; (b) the strength of the association between the covariate and the response (the risk factor potential); and (c) the association between the covariate and group (the extent of imbalance). In accordance with the developments in Section 4.4.3, they show that when (b) and (c) both exist, the unadjusted marginal odds ratio is positively biased, whereas the stratified-adjusted estimate is not. However, a strong association in both (b) and (c) must be present for a substantial bias to exist. They also show that heterogeneity of the odds ratios alone has little effect on the difference in power between the unadjusted versus adjusted analysis.

Since the marginal analysis is positively biased when (b) and (c) exist, then the marginal test appears to have greater power due to this bias and will misleadingly suggest that a smaller N is required. However, in this case the preferred analysis is the stratified analysis that is unbiased. However, a larger N is required to provide a given level of power.

4.12 POLYCHOTOMOUS AND ORDINAL DATA

The methods for the analysis of stratified 2×2 tables also apply to the analysis of polychotomous and ordinal data. Some of these have been implemented in the SAS PROC FREQ. Others are readily implemented using the basic principles described herein.

4.12.1 Cochran-Mantel-Haenszel Tests

The *cmh* option in SAS PROC FREQ provides three test statistics for an $R \times C$ table as described in Section 2.11. The test of general association on $(R - 1)(C - 1) df$ is computed using the conditional covariance matrix of a multinomial and thus equals $(N - 1)X_P^2/N$, X_P^2 being the Pearson contingency chi-square test in (2.143). The

test labeled "row mean scores differ" in SAS is the mean score test in (2.140), that with rank scores is asymptotically equivalent to the Wilcoxon test for two groups or the Kruskal-Wallis test for more than two groups. The test of nonzero correlation with rank scores is equivalent to the test of significance of the Spearman rank correlation. Landis et al. (1978) describe generalizations of these tests to the case of a stratified analysis. Computational details are presented by Stokes et al. (2000), among others. These stratified tests are based on the same principle employed by the Mantel-Haenszel test for stratified 2×2 tables, described in Section 4.2.3, where the differences between observed and expected frequencies are summed over strata. However, these stratified Mantel-Haenszel analyses do not provide stratified-adjusted estimates of the underlying parameters.

4.12.2 Stratified-Adjusted Estimates

Stratified-adjusted estimates of the magnitude of association between the classification variables aggregated over strata can be derived by application of weighted least squares as described in Section A.5. Consider the case where the frequencies among the C categories of a response variable are compared between $R = 2$ groups. The differences between groups can be described by the set of $C - 1$ odds ratios using one of the categories (say the first) as the reference as described in Section 2.10.3. In a stratified analysis, let $\hat{\boldsymbol{\theta}}_k$ denote the vector of $C - 1$ log odds ratios within the k th of K strata with estimated covariance matrix $\hat{\Sigma}_k$ obtained from (2.136). Under the assumption that $E(\hat{\boldsymbol{\theta}}_k) = \boldsymbol{\theta}$ for all strata, weighted least squares will provide a jointly minimum variance estimate of $\boldsymbol{\theta}$. In the framework of Section A.5.3, let $\mathbf{Y} = (\hat{\boldsymbol{\theta}}_1 // \dots // \hat{\boldsymbol{\theta}}_K)$ be the vector of $K(C - 1)$ odds ratios and $\mathbf{X} = (\mathbf{I}_{C-1} // \dots // \mathbf{I}_{C-1})$ be the $K(C - 1) \times (C - 1)$ design matrix of K concatenated identity matrices such that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$ with parameter vector $\boldsymbol{\theta}$. The errors are assumed to be distributed with mean zero and joint $K(C - 1) \times K(C - 1)$ covariance matrix $\hat{\Sigma} = \text{blockdiag}(\hat{\Sigma}_1 \dots \hat{\Sigma}_K)$. Then from (A.70) and (A.71) it is readily shown that

$$\hat{\boldsymbol{\theta}} = \left[\sum_{k=1}^K \hat{\Sigma}_k^{-1} \right]^{-1} \left[\sum_{k=1}^K \hat{\Sigma}_k^{-1} \hat{\boldsymbol{\theta}}_k \right], \quad (4.178)$$

with estimated covariance matrix

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \left[\sum_{k=1}^K \hat{\Sigma}_k^{-1} \right]^{-1}. \quad (4.179)$$

Example 4.26 Ethnicity and Hypertension in the Diabetes Prevention Program

Example 2.18 presents an analysis of the prevalence of hypertension as a function of the ethnicity of subjects enrolled into the Diabetes Prevention Program. The Pearson chi-square test of independence among the five ethnic categories was $X_P^2 = 41.5$ on 3 df . The following table presents the prevalence of hypertension stratified by

gender.

		Hypertension:			
		White	Afr. Am.	Hispanic	Other
Male	%	32.7%	40.4%	27.4%	37.0%
	n	730	193	201	119
Female	%	24.5%	35.4%	18.6%	19.4%
	n	1387	559	408	222

The unadjusted odds ratios and $\log(\widehat{OR})$ comparing each ethnicity versus Caucasians are

		Afr. Am.	Hispanic	Other
Male	OR	1.393	0.774	1.205
	$\widehat{\theta}_1 = \log(\widehat{OR})$	0.332	-0.256	0.187
Female	OR	1.676	0.705	0.740
	$\widehat{\theta}_2 = \log(\widehat{OR})$	0.516	-0.350	-0.301

The covariance matrix of the log odds ratio estimates within each stratum has elements

Male ($\widehat{\Sigma}_1$)		Female ($\widehat{\Sigma}_2$)	
Variance	Covariances	Variance	Covariances
Afr. Am.	0.02774	0.01173	
Hispanic	0.03125	0.00622	0.02007
Other	0.04228		0.00390
		0.03274	

where all covariances equal the value shown. Applying (4.178) and (4.179) yields the vector of joint minimum variance linear estimates and the corresponding covariance matrix of the estimates

$\widehat{\theta}$	$\widehat{V}(\widehat{\theta})$		
	Afr. Am.	Hispanic	Other
Afr. Am.	0.46942	0.00822	0.00239
Hispanic	-0.31317	0.00239	0.01222
Other	-0.08573	0.00236	0.00240
			0.01844

The corresponding odds ratios are 1.599, 0.731, and 0.918. If desired the significance of each pairwise group comparison could be assessed using a Wald test of the corresponding $\log(\widehat{OR})$ relative to its $S.E.$

While the above computations provide odds ratios relative to Caucasian, other adjusted odds ratios are also readily derived. For example, the log odds ratio comparing African American to Hispanic is the difference of the log odds ratios of each versus Caucasian [i.e., $0.46942 - (-0.31317) = 0.78259$], and the variance of the estimate is $0.00822 + 0.01222 - 2(0.00239) = 0.01566$.

The above computations were conducted using the program *MultMVLE DPP.sas*. The vectors of log odds ratios and their covariance matrices were computed by *stratlogits.sas*.

4.12.3 Vector Test of Homogeneity

This construction also provides a contrast (Cochran) test of homogeneity of the vectors of log odds ratios among strata using a contrast test as described in Section 4.6.1. For the above example, since $\mathbf{Y} = (\widehat{\theta}_1/\widehat{\theta}_2)$ contains two sets of three elements, the contrast matrix employed is

$$\mathbf{C}' = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}, \quad (4.180)$$

and the test on $(C - 1) = 3 \text{ df}$ is provided by

$$X_H^2 = \mathbf{Y}' \mathbf{C} (\mathbf{C}' \widehat{\Sigma} \mathbf{C})^{-1} \mathbf{C}' \mathbf{Y}, \quad (4.181)$$

where $\widehat{\Sigma} = \text{blockdiag}(\widehat{\Sigma}_1, \widehat{\Sigma}_2)$. For these data the resulting test is $X_H^2 = 4.97$ on 3 df with $p = 0.174$, so that the vectors of odds ratios are not significantly different between males and females. These computations are also provided by the program *MultMVLE DPP.sas*.

4.12.4 Stratified Mean Scores Estimate and Test

For a stratified mean scores test, the pivotal quantity is the group difference in the mean scores within each stratum using the mean table scores or the mean rank scores, among others. In the latter case, however, the rank score for a category within each stratum is proportional to the sample size for that stratum, in which case the mean fractional rank score should be employed. Within the k th stratum, the mean score difference \bar{d}_k and its estimated variance $\widehat{V}(\bar{d}_k)$ would be obtained from the numerator and denominator of (2.140). Then the MVLE mean score difference and its variance can be obtained by application of weighted least squares, as in Section 4.3.5. This estimate also provides a large sample stratified-adjusted test.

Example 4.27 Retinopathy in the DCCT

Example 2.17 compares the level of retinopathy progression between the DCCT intensive (*Int*) versus conventional (*Con*) groups after five years of follow-up among subjects who entered the study with preexisting nonsevere retinopathy. The minimal level of retinopathy at baseline was the presence of microaneurysms. The following table presents the levels of retinopathy progression (see Example 2.17) stratified by microaneurysms alone (*MA*) versus more severe lesions ($>\text{MA}$):

Stratum, Group	Retinopathy Level			
	0	1	2	3
<i>MA, Int</i>	220 (92.8%)	15 (6.3)	2 (0.8)	0
<i>MA, Con</i>	154 (81.1%)	28 (14.7)	6 (3.2)	2 (1.1)
$>\text{MA, Int}$	87 (82.1%)	4 (3.8)	10 (9.4)	5 (4.7)
$>\text{MA, Con}$	106 (76.8%)	5 (3.6)	12 (8.7)	15 (10.9)

The stratified-adjusted test could be based on table scores or fractional rank scores. Using the fractional ranks within each stratum (from the two groups combined), the mean fractional rank score (\bar{s}_{ik}) and its variance within each group ($i = 1, 2$) and stratum ($k = 1, 2$), and the difference between groups, $d_k = \bar{s}_{Int,k} - \bar{s}_{Con,k}$ and its variance are

	\bar{s}_{ik}	$\hat{V}(\bar{s}_{ik})$	d_k	$\hat{V}(d_k)$
MA, Int	1.033	1.434×10^{-4}	0.0596	2.584×10^{-4}
MA, Con	0.973	1.150×10^{-4}		
$>MA, Int$	1.013	3.042×10^{-4}	0.0305	7.003×10^{-4}
$>MA, Con$	0.983	3.961×10^{-4}		

The mean score difference between groups within each stratum then provides the basis for computation of a minimum variance linear estimate (MVLE) as described in Section 4.3.5. This yields a stratified-adjusted mean difference $d = 0.0518$ with $S.E. = 0.01374$ that yields a Wald test with $Z = 3.77$ and two-sided $p > 0.001$. A contrast test of homogeneity yields $p = 0.35$, indicating that the treatment group differences in the degree of retinopathy progression did not differ significantly between strata.

These computations were performed using the program *retinop strat.sas*.

4.12.5 Stratified Cochran-Armitage Test of Trend

Likewise, for a stratified Cochran-Armitage test of trend, the pivotal quantity within each stratum is the estimate of the slope using either table scores or fractional rank scores, among others, provided by (2.148). Then the stratified test can be based on the MVLE of the mean slope.

Example 4.28 Coronary Heart Disease in the Framingham Study

Example 2.21 presents the prevalence of CHD within cholesterol level categories and Example 2.19 presents the numbers of subjects within each category of systolic blood pressure (SBP) and cholesterol. The following table presents the numbers and percent with CHD within each cholesterol category stratified by the category of blood pressure, the corresponding estimate of the slope using fractional rank scores, and the corresponding *S.E.*

SBP (mmHg)	Cholesterol (mg/dL)				$\hat{\beta}_{f,k}$	<i>S.E.</i>
	<200	200–219	220–259	260+		
<127	2 (1.7%)	3 (3.4)	8 (6.3)	7 (9.5)	0.10002	0.03842
127–146	3 (2.4%)	2 (2.0)	11 (5.0)	12 (10.8)	0.10381	0.03369
147–166	3 (6.0%)	0	6 (8.1)	11 (19.3)	0.19701	0.06847
167+	4 (15.4)	3 (13.0)	6 (12.2)	11 (25.0)	0.14347	0.11366

The MVLE stratified-adjusted estimate of the slope using fractional ranks is $\hat{\beta}_f = 0.115$ with $S.E. = 0.0233$, that yields a Wald stratified-adjusted test of trend with $Z = 4.938$ and $p < 0.0001$ (two-sided). Using a contrast test of homogeneity yields

$X_H^2 = 1.76$ on 3 df with $p = 0.63$. Thus, the association of cholesterol with risk of CHD does not vary significantly over the levels of blood pressure, and the slope for trend remains highly significant.

These computations were performed using *MVLE trend.sas*.

4.13 PROBLEMS

4.1 Show that Cochran's test for multiple 2×2 tables is equal to

$$X_{U,C}^2 = \frac{(\sum_k [a_k - E(a_k)])^2}{\sum_k \hat{V}_u(a_k)} = \frac{[\sum_k \hat{w}_k (p_{1k} - p_{2k})]^2}{\sum_k \hat{w}_k^2 \hat{V}_{0k}}, \quad (4.182)$$

with

$$\hat{w}_k = \frac{1}{\left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}}\right)} \quad \hat{V}_{0k} = p_k (1 - p_k) \left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}}\right). \quad (4.183)$$

4.2 Assume that there is a common value of θ for all strata, or $H_0: \theta_k = \theta$ such that $E(\hat{\theta}_k) = \theta$ for all k . Thus, we wish to combine the estimates over strata to estimate θ .

4.2.1. Under a fixed effects model (4.30) and (4.31), use weighted least squares (see Section A.5) to show that

$$\hat{\theta} = \sum_k \hat{\tau}_k \hat{\theta}_k / \sum_\ell \hat{\tau}_\ell \quad (4.184)$$

and

$$\hat{V}(\hat{\theta}) = \left(\sum_k \hat{\tau}_k\right)^{-1}, \quad (4.185)$$

where $\hat{\tau}_k = \hat{V}(\hat{\theta}_k)^{-1} = \hat{\sigma}_k^{-2}$.

4.2.2. Also show that the variance of the estimate can be obtained as

$$V(\hat{\theta}) = \sum_k \omega_k^2 \sigma_k^2 = \frac{1}{\sum_k \sigma_k^{-2}}. \quad (4.186)$$

4.2.3. For the risk difference, log odds ratio, and log relative risk, use the above to obtain the expressions for $\hat{\tau}_k$, the weights \hat{w}_k , the pooled *MVLE* $\hat{\theta}$, $V(\hat{\theta})$, and a consistent estimate $\hat{V}(\hat{\theta})$. In these derivations use the expressions for the variance of $\hat{\theta}_k$ under the assumption that a difference exists ($\pi_{1k} \neq \pi_{2k}$).

4.2.4. Do likewise for θ_k defined as $\theta_k = g(\pi_{1k}) - g(\pi_{2k})$ for these scales:

1. The arcsin transformation $g(\pi) = \sin^{-1}(\sqrt{\pi})$.
2. The probit transformation $g(\pi) = \Phi^{-1}(\pi)$, where $\Phi(\cdot)$ is the standard normal cdf.
3. The square root transformation $g(\pi) = \sqrt{\pi}$.

4.3 The Mantel-Haenszel estimate of the common odds ratio is also a linear estimator, as shown in (4.15) and (4.16).

4.3.1. Show that these \widehat{v}_k are not proportional to $V(\widehat{OR}_k)^{-1}$, where $V(OR_k)$ was derived in Problem 2.5.4. Thus, the variance of the Mantel-Haenszel estimate \widehat{OR}_{MH} is greater than that of an MVLE efficient estimate of the common odds ratio.

4.3.2. Also, show that for $OR_k \doteq 1$, the \widehat{OR}_{MH} is approximately the MVLE of the common odds ratio (not the common log odds ratio).

4.4 The Radhakrishna family of tests $X_{A(g)}^2$ is presented in (4.97) using estimated optimal weights $\{\widehat{w}_k\}$ in (4.96) specific to the desired scale $\theta = g(\pi_1) - g(\pi_2)$, and where $X_{A(g)}^2$ is asymptotically distributed as chi-square on 1 df under H_0 : $\pi_{1k} = \pi_{2k} = \pi_k \forall k$.

4.4.1. Adopt a local alternative of the form $\pi_{1k} = \pi_k + \delta n^{-\frac{1}{2}}$ and $\pi_{2k} = \pi_k - \delta n^{-\frac{1}{2}}$. Use a Taylor's expansion of $g(\pi_{ik})$ about $\pi_k \in (\pi_{1k}, \pi_{2k})$ for each group $i = 1, 2$ to derive (4.82) and (4.83).

4.4.2. Using the expression for the optimal weights that maximize the asymptotic efficiency, $\mathbf{W} = \Sigma_0^{-1} \mathbf{G}$ in (4.94), show that the optimal weights for each of the following scales are estimated as:

Scale	\widehat{w}_k
Risk difference	$\left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right)^{-1} \frac{1}{p_k(1-p_k)}$
Log relative risk	$\left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right)^{-1} \frac{1}{1-p_k}$
Log odds ratio	$\left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right)^{-1}$

that simplify to the expressions in Table 4.8.

4.4.3. Derive the explicit expression for the test for each scale presented below.

1. Constant risk difference:

$$X_{A(RD)}^2 = \frac{\left[\sum_k \widehat{V}_{0k}^{-1} (p_{1k} - p_{2k}) \right]^2}{\sum_k \widehat{V}_{0k}^{-1}} \quad (4.187)$$

2. Constant log relative risk:

$$X_{A(RR)}^2 = \frac{\left[\sum_k p_k \widehat{V}_{0k}^{-1} (p_{1k} - p_{2k}) \right]^2}{\sum_k p_k^2 \widehat{V}_{0k}^{-1}} \quad (4.188)$$

3. Constant log odds ratio:

$$X_{A(OR)}^2 = \frac{\left[\sum_k p_k (1-p_k) \widehat{V}_{0k}^{-1} (p_{1k} - p_{2k}) \right]^2}{\sum_k [p_k (1-p_k)]^2 \widehat{V}_{0k}^{-1}} \quad (4.189)$$

4.4.4. The additional scales used in Problem 4.2.4 are also members of the Radhakrishna family. Show that the estimated optimal weights for the test for a common difference on each of these scales are as follows:

Scale	\widehat{w}_k
Arcsin	$\frac{n_{1k}n_{2k}}{N_k} \sqrt{\frac{1}{p_k(1-p_k)}}$
Probit	$\left(\frac{n_{1k}n_{2k}}{N_k}\right) \frac{\exp[-\Phi^{-1}(p_k)^2/2]}{p_k(1-p_k)\sqrt{2\pi}}$
Square root	$\left(\frac{n_{1k}n_{2k}}{N_k}\right) \frac{1}{(1-p_k)\sqrt{p_k}}$

4.5 For a scale with measure $\theta_k = g(\pi_{1k}) - g(\pi_{2k})$, we now show that the Radhakrishna test of association may also be viewed as an estimation-based test using an MVLE of the common parameter, but using weights estimated under H_0 rather than under H_1 .

4.5.1. Since $\widehat{\theta}_k \cong g'(p_k)(p_{1k} - p_{2k})$ show that the numerator of $X_{A(g)}^2$ in (4.97) is approximately a weighted average of the $\{\widehat{\theta}_k\}$ of the form

$$T_g \cong \sum_k \frac{\widehat{\theta}_k}{g'(\pi_k)^2 V_{0k}}. \quad (4.190)$$

4.5.2. From the δ -method show that asymptotically

$$V(T_g | H_0) \cong \sum_k \frac{1}{g'(\pi_k)^2 V_{0k}}. \quad (4.191)$$

4.5.3. Thus, asymptotically the test can be expressed as

$$X_{A(g)}^2 \cong \frac{\left[\sum_k \omega_{0k} \widehat{\theta}_k \right]^2}{\sum_k \omega_{0k}^2 V(\widehat{\theta}_k | H_0)} \quad (4.192)$$

with weights

$$\omega_{0k} = V(\widehat{\theta}_k | H_0)^{-1} \cong \left[g'(p_k)^2 \widehat{V}_{0k} \right]^{-1}. \quad (4.193)$$

4.6 Now let $Z_{A(g)}$ be the standardized normal deviate test corresponding to the Radhakrishna test $X_{A(g)}^2$ in (4.97). Consider two separate tests of association on two different scales, such as Z_r for the test of common log odds ratio and Z_s for the test of common log relative risk.

4.6.1. Show that the correlation among the two standardized deviates is

$$\text{corr}(Z_r, Z_s) = \rho_{rs} = \frac{\sum_k w_{rk} w_{sk} \pi_k (1 - \pi_k) \left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right)}{(V_r V_s)^{1/2}}, \quad (4.194)$$

where V_r and V_s are the denominators in the Radhakrishna test X_A^2 for the r th and s th scales, respectively.

Hint: For a vector \mathbf{X} with covariance matrix Σ_x , the covariance of two bilinear forms $\mathbf{A}'\mathbf{X}$ and $\mathbf{B}'\mathbf{X} = \mathbf{A}'\Sigma_x\mathbf{B}$.

4.6.2. Consider two scales $g_s(\pi)$ and $g_r(\pi)$, where $\theta_{sk} = \theta_s \forall k$ for scale $g_s(\pi)$. Thus, the test T_s is optimal and test T_r is not. Show that the $\text{ARE}(T_r, T_s | \theta_s) = \rho_{rs}^2$, as presented in (4.116).

4.6.3. Let Z_r and Z_s be the standardized deviates corresponding to T_r and T_s . Show that $\text{ARE}(T_r, T_s | \theta_s) = \text{ARE}(Z_r, Z_s | \theta_s)$.

4.6.4. Using the sample fractions described in (4.118), show that the sample size N factors from the above expression for ρ_{rs}^2 using weights $\tilde{w}_{\ell k}$ presented in (4.119) to yield

$$\text{ARE}(T_r, T_s | \theta_s) = \frac{\left[\sum_k \tilde{w}_{rk} \tilde{w}_{sk} \phi_{0k}^2 \right]^2}{\left[\sum_k \tilde{w}_{rk}^2 \phi_{0k}^2 \right] \left[\sum_k \tilde{w}_{sk}^2 \phi_{0k}^2 \right]} = \rho_{rs}^2. \quad (4.195)$$

4.6.5. To compute the $\widehat{\text{ARE}}$ for each pair of scales, the covariance of the two tests involves the cross products $\widehat{w}_{rk} \widehat{w}_{sk} \widehat{V}_{0k}$ that can be simplified. Let

$$M_k = \left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right)^{-1} = \frac{n_{1k} n_{2k}}{N_k} \quad (4.196)$$

so that $\widehat{V}_{0k} = p_k(1 - p_k)/M_k$. Then for the following pairs of scales, show that these terms reduce as follows:

Pair of Tests	$\widehat{w}_{rk} \widehat{w}_{sk} \widehat{V}_{0k}$
log Odds ratio, Risk difference	M_k
log Odds ratio, log Relative risk	$p_k M_k$
log Relative risk, Risk difference	$M_k / (1 - p_k)$

4.7 For the measurement error random effects model of Section 4.10.1, assume that the random error variances vary from subject to subject such that $V(\varepsilon_i) = \sigma_{\varepsilon_i}^2$ is not constant for all subjects in the population, such as where $\sigma_{\varepsilon_i}^2$ is a function of the conditional mean $E(y|v_i)$. Also assume that a consistent estimate of the error variance $\widehat{\sigma}_{\varepsilon_i}^2$ is provided for each subject.

4.7.1. Using the partitioning of variation as in (A.6), show that $V(y_i) = \sigma_{\varepsilon_i}^2 + \sigma_v^2$ and that $V(Y) = E(\sigma_{\varepsilon_i}^2) + \sigma_v^2$.

4.7.2. For a sample of N independent observations, show that

$$V \left[\sum_i y_i \right] = NV(Y) = E \left[\sum_i \sigma_{\varepsilon_i}^2 \right] + N\sigma_v^2. \quad (4.197)$$

4.7.3. Then show that the moment estimator for the mixing distribution variance σ_v^2 is provided by

$$\hat{\sigma}_v^2 = \hat{V}(Y) - \sum_i \hat{\sigma}_{\varepsilon_i}^2 / N, \quad (4.198)$$

where

$$\hat{V}(Y) = S_y^2 = \frac{\sum_i (y_i - \bar{y})^2}{N - 1}. \quad (4.199)$$

4.7.4. Alternatively, partition the sum of squares of Y about μ as in (A.4) to show that

$$S_y^2 = \frac{\sum_i (y_i - \mu)^2 - N(\bar{y} - \mu)^2}{N - 1} \quad (4.200)$$

and that

$$E(S_y^2) = \frac{\sum_i \sigma_{\varepsilon_i}^2}{N} + \sigma_v^2. \quad (4.201)$$

4.7.5. Then again show that the moment estimator for σ_v^2 is provided by (4.198).

4.8 For the DerSimonian-Laird random effects model in Section 4.10.2, do the following:

4.8.1. Show that (4.148) can be expressed as

$$E(X_{H,C}^2) = \sum_k \tau_k \left(\sigma_{\hat{\theta}_k}^2 + \sigma_{\theta}^2 \right) - \left(\sum_k \tau_k \right) V(\hat{\mu}_{\theta}). \quad (4.202)$$

4.8.2. Then derive (4.150).

4.8.3. Then derive (4.152), (4.153), and (4.154).

4.8.4. In a fixed-point iterative procedure as described in (4.156)–(4.158), at the first step, show the expressions for the weights $\hat{\omega}_k^{(1)}$ and the estimated variance $\hat{V}(\hat{\mu}_{\theta}^{(1)})$ for the log odds ratio.

4.9 A cross-sectional occupational health study assessed the association between smoking and the prevalence of byssinosis among workers in a textile plant (Higgins and Koch, 1977). The following 2×2 tables were obtained giving the frequencies of workers with byssinosis versus not among smokers and nonsmokers stratified by the length of employment:

Stratum	Smoker		Nonsmoker	
	a_k	n_{1k}	b_k	n_{2k}
1: <10 years	44	1587	19	1142
2: 10–20 years	21	481	5	231
3: >20 years	60	1121	16	857

(reproduced with permission). The following analyses assess the increase in risk, if any, associated with smoking versus not smoking.

4.9.1. Separately within each strata, and for the data combined into a single marginal 2×2 table with total $N = 5419$, calculate the following:

1. V_u and the contingency (Pearson) chi-square statistic and p -value.
2. V_c and the Mantel chi-square statistic and p -value.
3. The difference in proportions, the log odds ratio, and the log relative risk, and the estimated variance of each.
4. The 95% confidence limits for the difference, odds ratio, and relative risk.

4.9.2. Now conduct a stratified Mantel-Haenszel analysis of odds ratios.

1. Calculate the Mantel-Haenszel estimate of the common odds ratio and the Robins et al. variance of the log odds ratio. Use these to compute asymmetric 95% confidence limits on the odds ratio.
2. Calculate the Mantel-Haenszel and Cochran stratified-adjusted tests of significance.

3. Now use the Mantel-Haenszel test value to obtain the test-inverted confidence limits for the common odds ratio.

4. Compare and comment on the difference between the marginal unadjusted analysis and the stratified-adjusted analysis of the odds ratios.

4.9.3. Now conduct a stratified Mantel-Haenszel analysis of relative risks.

1. Calculate the Mantel-Haenszel estimate of the common relative risk.
2. Now use the Mantel-Haenszel test value to obtain the test-inverted confidence limits for the common relative risk.
3. Compare and comment on the difference between the marginal unadjusted analysis and the stratified-adjusted analysis of the relative risks.

4.9.4. For each of the three scales (difference, log odds ratio, log relative risk) perform the following *MVLE* analyses over all three strata.

1. Obtain the estimate of the common parameter and its variance.
2. Calculate the 95% confidence interval on the parameter under each scale, and also the asymmetric confidence limits for the odds ratio and relative risk.
3. Compare the weights for the estimates of the common parameter on each scale.
4. Compare the *MVLE* estimates and confidence limits to the marginal unadjusted analysis.
5. Compute the relative weights (summing to 1) for the Mantel-Haenszel estimates of the common odds ratio and relative risk and compare these to the weights for the *MVLE* estimates for the log odds ratio and log relative risk, respectively.
6. Then compare the stratified-adjusted estimates of the odds ratio and the relative risk and their confidence limits.

4.9.5. Evaluate the extent to which years of employment strata are associated with the smoking group and/or are a risk factor for response (byssinosis) using the appropriate 3×2 tables. What are the implications for the comparison of the adjusted versus unadjusted analyses?

4.9.6. Conduct the following tests of significance.

1. The 3 df omnibus X^2 test using the risk difference ($p_{1k} - p_{2k}$) within each stratum. Also state the null and alternative hypotheses and the result of the test.

2. An alternative is to use three separate tests (Cochran or Mantel-Haenszel), one within each stratum, at $\alpha = 0.05/3$. Interpret the tests conducted under Problems 4.9.1.1 and 4.9.1.2 above in terms of these Bonferroni-adjusted significance levels.

3. Conduct Cochran's $K-1$ *df* test of homogeneity for each of the three scales.

4. For each of the three scales, perform the Radhakrishna test of association.

Compare the tests on each scale in relation to the tests for homogeneity.

5. Compare the Mantel-Haenszel test obtained in Problem 4.9.2.2 to the Radhakrishna test for log odds.

6. Compute the correlation $\hat{\rho}_{rs}$ and $\widehat{ARE}(T_r, T_s)$ for each of the three possible pairs of scales among the risk difference, log relative risk, and log odds ratio.

7. Confirm that the Gastwirth *MERT* conditions apply and then compute the scale robust *MERT* in (4.125). Compare this test to the individual Radhakrishna family tests. Describe the *ARE* of this test relative to others in the family of tests considered.

8. Likewise, compute the Wei-Lachin test of stochastic ordering using the risk difference within each stratum. State the null and alternative hypotheses. Compare to the other tests above.

4.9.7. Using a random effects model for an analysis of log odds ratios:

1. Estimate $\hat{\sigma}_\theta^2$.

2. Compute the updated $V(\hat{\theta}_k)$.

3. Compute the $\hat{\omega}_k^{(1)}$.

4. Compute $\hat{\mu}_\theta^{(1)}$.

5. Compute $V(\hat{\mu}_\theta^{(1)})$.

6. Compute asymmetric 95% confidence limits on θ .

7. Compare these computations under a random effects model to the previous computations under a fixed effects model.

4.10 Using the developments in Section 4.11.1:

4.10.1. Derive the expression for $\sigma_{\bar{T}_0}^2$ in (4.162).

4.10.2. Derive the expression for $\mu_{\bar{T}}$ in (4.163).

4.10.3. Derive the expression for $\sigma_{\bar{T}_0}^2$ in (4.165).

4.10.4. Derive the expression in (4.168).

4.11 Consider an epidemiologic study designed to compare the risks (odds) of exposure to a risk factor (E vs. \bar{E}) on the incidence of developing a disease over a fixed period of time (D vs. \bar{D}). Because the study will not be randomized, it is planned to conduct a stratified analysis with the following expected sample fractions and control (\bar{E}) group probabilities, and the specified odds ratios within each of three strata:

Stratum	Stratum		Group			
	Fraction	Fractions	ξ_{1k}	ξ_{2k}	π_{2k}	OR_k (e^{θ_k})
Stratum	ζ_k					
1	0.15		0.3	0.7	0.30	1.80
2	0.50		0.6	0.4	0.22	2.50
3	0.35		0.45	0.55	0.15	3.00

- 4.11.1.** Compute the corresponding probabilities in the exposed group π_{1k} .
- 4.11.2.** Compute the marginal probabilities π_{1*} and π_{2*} obtained as the weighted average of the π_{1k} and π_{2k} weighting by the expected fraction within each stratum. From these compute the sample size required to provide power of 0.90 for the unadjusted test of the difference between two proportions in the unstratified analysis with $\alpha = 0.05$, two-sided.
- 4.11.3.** Now compute the sample size required for the stratified analysis of a common odds ratio to provide power of 0.90.
- 4.11.4.** Suppose that the study is conducted with a total sample size of $N = 150$. What level of power is provided by the stratified analysis?
- 4.11.5.** Alternatively, for these expected probabilities within each stratum, compute the sample size required to provide power of 0.90 for the test of a common relative risk rather than a common odds ratio.
- 4.11.6.** Explain why the required sample sizes for the test of a common odds ratio differs from that for the test of a common relative risk.
- 4.12** Byar (1985) presents data from a Veterans Administration Cooperative Urologic Research Group (VACURG) prostate cancer study. The data are provided in *vacurg85.dat* on this book's website and are also used in Problem 9.17. Three baseline variables of interest are performance status (*pf*) coded as normal (0), < 50% time in bed (1), 50 to < 100% time in bed (2), and confined to bed (3); EKG findings (*ekg*) coded as normal (0), benign (1), rhythmic disturbances and electrolyte changes (2), heart blocks or conduction defects (3), heart strain (4), old myocardial infarct (MI) (5), recent MI (6); and history of cardiovascular disease (*hx*) coded as no (0), yes (1). Owing to sparseness of the data, the program *vacurgbase.sas* defines two new variables: *inbed* is 0 if *pf* = 0 and 1 if bedridden (*pf* \geq 1); and *ekgc* is constructed by collapsing the categories (0,1), (2,3), and (5,6) of *ekg*. Use SAS PROC FREQ to assess the association between being bedridden (*inbed* = 1) and the collapsed EKG status (*ekgc*) using the test of general association within the full cohort, those with and without a history of cardiovascular disease, and stratified by history of CVD. Likewise, in the full cohort and then within strata, compute the generalized odds ratios of being bedridden among EKG categories using 0 as the reference. Compute the joint minimum variance estimate of the combined vector of odds ratios and their *S.E.s*, and the vector test of homogeneity of the generalized odds ratios among the two strata. Interpret the differences between the unstratified and stratified analysis results.
- 4.12.1.** Similarly, using rank scores, compute the Cochran-Armitage trend test within the full cohort, the two CVD history strata, and adjusted over strata; and within each compute the estimated coefficient using fractional ranks ($\hat{\beta}_f$) and the associated *S.E.*. Then compute the MVLE of the common coefficient and its *S.E.* and conduct a test of homogeneity. Interpret the differences among the results.

Case-Control and Matched Studies

The preceding chapters considered exclusively statistical methods for the assessment of the association between two or more independent treatment or exposure groups and the risk of a binary or polychotomous outcome from a cross-sectional or prospective study. In this chapter we consider two generalizations. One is the case-control or retrospective study. When separate independent groups of cases and controls are constructed, many of the methods in the preceding chapters still apply. However, the odds ratio and relative risk have different interpretations in a retrospective study from those in a prospective study.

The other extension is to the matched study. Matching is often used in a case-control study to control for important factors; however, matching may be used in a prospective study as well. In either case the two exposure (or treatment) groups, and the case-control groups, are not independent, and thus different statistical methods must be employed.

5.1 UNMATCHED CASE-CONTROL (RETROSPECTIVE) SAMPLING

Consider an unmatched case-control or retrospective study in which a sample of known cases is obtained who have the outcome disease or characteristic of interest (D) and who are to be compared with an independent sample of nondiseased controls (\bar{D}). For each subject in the two groups, the prior degree of exposure to the risk factor under study, classified as E and \bar{E} , is then determined *retrospectively*. Just the opposite is done in a prospective study, where samples of exposed and nonexposed individuals (E and \bar{E}) are obtained and then followed prospectively to determine whether each individual develops the disease (D) or does not (\bar{D}). The entries in a

2×2 table of frequencies from each type of study are

Prospective		Retrospective	
Group		Group	
Exposed	Not Exposed	Cases	Controls
E	\bar{E}	D	\bar{D}
D	a	b	m_1
\bar{D}	c	d	m_2
n_1	n_2	n_1	n_2
	N		N

(5.1)

In each case the "unconditional" likelihood given samples of sizes n_1 and n_2 is a product binomial likelihood, as shown in (2.53); and conditioning on all margins fixed, the conditional likelihood is a hypergeometric probability as shown in (2.60). Therefore, Fisher's exact test or the large sample contingency (Cochran) or Mantel-Haenszel 1 df χ^2 tests can be used with such data to test the hypothesis of association between disease (case-control) and exposure. However, there are additional considerations in describing the degree of association.

5.1.1 Odds Ratio

Since the data are obtained from retrospective sampling, the retrospective odds ratio can be distinguished from the prospective odds ratio. Given retrospective samples of n_1 cases (D) and n_2 controls (\bar{D}), the assumed conditional probabilities are

		D	\bar{D}	
		E	\bar{E}	
ϕ_1	ϕ_2			
$1 - \phi_1$	$1 - \phi_2$			
1.0	1.0			

(5.2)

where ϕ_1 and ϕ_2 are the retrospective *disease conditional probabilities* of exposure (E) given being a case (D) versus a control (\bar{D}):

$$\phi_1 = P(E|D) \text{ and } \phi_2 = P(E|\bar{D}). \quad (5.3)$$

Thus, the *retrospective odds ratio* of exposure given disease is

$$OR_{retro} = \frac{\phi_1 / (1 - \phi_1)}{\phi_2 / (1 - \phi_2)} = \frac{P(E|D) / P(\bar{E}|D)}{P(E|\bar{D}) / P(\bar{E}|\bar{D})} \hat{=} \frac{ad}{bc}, \quad (5.4)$$

where $\hat{=}$ means "estimated as" using the cell frequencies from the retrospective 2×2 table.

However, it is of far greater interest to describe the *prospective odds ratio* (OR) of disease given exposure as presented in Chapter 2,

$$OR = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = \frac{P(D|E) / P(\bar{D}|E)}{P(D|\bar{E}) / P(\bar{D}|\bar{E})}. \quad (5.5)$$

To obtain this quantity we must also account for the prevalence of the disease in the population from which the cases and controls arose:

$$P(D) = \delta. \quad (5.6)$$

In this case the joint and marginal probabilities of disease and exposure become

		Cases	Controls	
		D	\bar{D}	
E	$P(E, D)$	$P(E, \bar{D})$	$P(E)$	
	$P(\bar{E}, D)$	$P(\bar{E}, \bar{D})$	$P(\bar{E})$	
		δ	$1 - \delta$	1.0

where, for example, $P(E, D) = P(E|D)P(D) = \phi_1\delta$, and so forth. Therefore, these probabilities are expressed as

		Cases	Controls	
		D	\bar{D}	
E	$\phi_1\delta$	$\phi_2(1 - \delta)$	$\phi_1\delta + \phi_2(1 - \delta)$	
	$(1 - \phi_1)\delta$	$(1 - \phi_2)(1 - \delta)$	$(1 - \phi_1)\delta + (1 - \phi_2)(1 - \delta)$	
		δ	$1 - \delta$	1.0

The prospective *exposure-conditional probabilities* of disease given exposure status, $\pi_1 = P(D|E)$ and $\pi_2 = P(D|\bar{E})$, can then be obtained as the ratio of the joint to the marginal probabilities, such as $P(D|E) = P(D, E)/P(E)$, which is equivalent to using Bayes' theorem. Thus, these prospective probabilities can then be expressed in terms of the disease-conditional probabilities (ϕ_1, ϕ_2) and the prevalence δ as

$$\begin{aligned} \pi_1 &= P(D|E) = \frac{P(D, E)}{P(E)} = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|\bar{D})P(\bar{D})} \quad (5.9) \\ &= \frac{\phi_1\delta}{\phi_1\delta + \phi_2(1 - \delta)} \end{aligned}$$

and

$$\pi_2 = P(D|\bar{E}) = \frac{(1 - \phi_1)\delta}{(1 - \phi_1)\delta + (1 - \phi_2)(1 - \delta)}. \quad (5.10)$$

It then follows that the prospective odds ratio of disease given exposure presented in (5.4) equals the retrospective odds ratio of exposure given disease, so that

$$\begin{aligned} OR &= \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = \frac{P(D|E) / P(\bar{D}|E)}{P(D|\bar{E}) / P(\bar{D}|\bar{E})} \quad (5.11) \\ &= \frac{P(E|D) / P(\bar{E}|D)}{P(E|\bar{D}) / P(\bar{E}|\bar{D})} = \frac{\phi_1 / (1 - \phi_1)}{\phi_2 / (1 - \phi_2)} = OR_{retro}. \end{aligned}$$

Even though the prevalence (probability) of the disease δ is required to estimate the prospective exposure conditional probabilities, these terms cancel from the expression for the odds ratio.

Thus, all of the methods described previously for the analysis of odds ratios in single and multiple 2×2 tables for two independent groups from a cross-sectional or prospective study also apply to the analysis of a case-control study with independent groups. The only alteration is that the disease conditional probabilities (ϕ_1, ϕ_2) are substituted for the exposure conditional probabilities (π_1, π_2) in the expressions for the underlying parameters. The estimates and tests are then obtained using the sample proportions from the retrospective 2×2 table in (5.1) computed as $p_1 = a/n_1$, $p_2 = b/n_2$, and $p = m_1/N$, where $p_1 = \hat{\phi}_1$, $p_2 = \hat{\phi}_2$, and $p = \hat{\phi}$.

5.1.2 Relative Risk

Now consider the relative risk. Based on a case-control study, the *retrospective relative risk* of exposure given disease versus no disease is

$$RR_{retro} = \frac{\phi_1}{\phi_2} = \frac{P(E|D)}{P(E|\bar{D})} \hat{=} \frac{\hat{\phi}_1}{\hat{\phi}_2} = \frac{a/n_1}{b/n_2}. \quad (5.12)$$

However, the parameter of interest is the *prospective* relative risk of disease given exposure versus no exposure, that is defined as

$$RR = \frac{\pi_1}{\pi_2} = \frac{P(D|E)}{P(D|\bar{E})} \neq RR_{retro}. \quad (5.13)$$

Thus, the prospective relative risk cannot be obtained from the retrospective relative risk.

However, through application of Bayes' theorem, or from (5.8), Cornfield (1951) showed that the prospective probabilities π_1 and π_2 can be defined directly given knowledge (specification) of the prevalence of the disease δ . From the resulting expressions for π_1 and π_2 in (5.9) and (5.10), the prospective relative risk is

$$\begin{aligned} RR &= \frac{P(D|E)}{P(D|\bar{E})} = \left[\frac{\delta\phi_1}{\delta\phi_1 + (1 - \delta)\phi_2} \right] \left[\frac{\delta(1 - \phi_1) + (1 - \delta)(1 - \phi_2)}{\delta(1 - \phi_1)} \right] \\ &= \left[\frac{\phi_1}{1 - \phi_1} \right] \left[\frac{\delta(1 - \phi_1) + (1 - \delta)(1 - \phi_2)}{\delta\phi_1 + (1 - \delta)\phi_2} \right]. \end{aligned} \quad (5.14)$$

Given knowledge of the prevalence δ , this quantity can be estimated consistently as

$$\widehat{RR} = \frac{\widehat{\pi}_1}{\widehat{\pi}_2} = \left[\frac{\delta p_1}{\delta p_1 + (1 - \delta)p_2} \right] \left[\frac{\delta(1 - p_1) + (1 - \delta)(1 - p_2)}{\delta(1 - p_1)} \right]. \quad (5.15)$$

Since the prevalence of the disease is often unknown, and is usually not directly estimable from such a study, this approach is rarely practical.

Cornfield (1951), however, showed that when the disease is rare ($\delta \downarrow 0$), the value of the RR is approximately

$$RR_{\delta \downarrow 0} \doteq \left(\frac{\phi_1}{1 - \phi_1} \right) \left(\frac{1 - \phi_2}{\phi_2} \right) = OR_{retro}, \quad (5.16)$$

which is the retrospective odds ratio of exposure given disease. Thus, under the assumption that the prevalence of the disease is rare ($\delta \doteq 0$), the retrospective odds ratio provides an approximate estimate of the prospective relative risk.

It can also be shown that an equivalent condition is that $\delta(\phi_1 - \phi_2) \downarrow 0$, or

$$RR_{\delta(\phi_1 - \phi_2) \downarrow 0} \doteq OR_{retro}, \quad (5.17)$$

so that the approximation applies when either the prevalence becomes rare, or the difference $(\phi_1 - \phi_2)$ approaches the null hypothesis.

5.1.3 Attributable Risk

As for the precise definition of the relative risk in (5.14), other measures of association, such as the risk difference or attributable risk, require that explicit expressions be derived using the known or specified value of the probability of the disease (δ) in conjunction with Bayes' Theorem. As a result, the analysis of a case-control study usually focuses only on the estimation and description of the odds ratio. The odds ratio applies to any case-control study because the retrospective odds ratio equals the prospective odds ratio. In the case of a rare disease, the odds ratio can also be interpreted approximately as a relative risk.

Since the attributable risk and the population attributable risk are also a function of the relative risk, then under the rare disease assumption the odds ratio also provides an estimate of the measures of attributable risk. As described in Section 2.8.2 for a prospective study, since the population attributable risk is a direct function of the relative risk, then the point estimate and asymmetric confidence limits can be obtained from the point estimate and asymmetric confidence limits for the odds ratio.

From (2.110), under the rare disease assumption as in (5.16), the population attributable risk may be estimated as a function of the odds ratio as

$$\widehat{PAR} \doteq \frac{\alpha_1(\widehat{OR} - 1)}{\alpha_1\widehat{OR} + \alpha_2}, \quad (5.18)$$

where the prevalence of exposure $\alpha_1 = P(E)$ is assumed to be known or specified. Using the δ -method (see Problem 5.1), it is then readily shown that asymptotically

$$V(\text{logit } \widehat{PAR}) \cong \left[\frac{OR}{OR - 1} \right]^2 V[\log OR], \quad (5.19)$$

where the exposure prevalence α_1 cancels from this expression. This variance is consistently estimated by substituting the retrospective estimate \widehat{OR} and the estimate of the variance $\widehat{V}[\log \widehat{OR}]$ from (2.47).

Example 5.1 *Smoking and Lung Cancer*

Cornfield (1951) presents the following example to describe the accuracy of his approximation. Schrek et al. (1950) reported a case-control study of smoking (E) and lung cancer (D). The retrospectively sampled data are

	D	\bar{D}
E	154	80
\bar{E}	46	120
	200	200

where $p_1 = \hat{\phi}_1 = 0.77$, $p_2 = \hat{\phi}_2 = 0.40$, and $\widehat{OR}_{retro} = 5.0217$. Under the rare disease assumption, this provides an approximation to the prospective relative risk.

How close is this to an estimate of the relative risk? From a large population-based study, Dorn (1944) reported that the prevalence of lung cancer at that time was 15.5/100,000 population. Thus, $\delta = P(D) = 0.155 \times 10^{-3}$. The joint and marginal probabilities of the exposure and disease categories are estimated to be

	D	\bar{D}	
E	0.119 $\times 10^{-3}$	0.399938	0.400057
\bar{E}	0.036 $\times 10^{-3}$	0.599907	0.599943
	0.155 $\times 10^{-3}$	0.999845	1.0

so that $\hat{\pi}_1 = \hat{P}(D|E) = 0.000119/0.400057 = 0.297 \times 10^{-3}$ and $\hat{\pi}_2 = \hat{P}(D|\bar{E}) = 0.000036/0.599943 = 0.060 \times 10^{-3}$. Therefore, the direct estimate of the prospective relative risk given the above value of δ is $\widehat{RR} = \hat{\pi}_1/\hat{\pi}_2 = 0.297/0.060 = 4.95$. The odds ratio of 5.02 provides a relatively accurate estimate of this relative risk. The asymmetric 95% confidence limits are (3.25, 7.75) and the association is highly significant with a Mantel-Haenszel test value of 56.25 on 1 df , $p < 0.001$.

To estimate the population attributable risk, the prevalence of exposure to smoking in the population must be specified. For $\alpha_1 = 0.3$ the estimate is $\widehat{PAR} = 0.54680$, which indicates that smoking, if the sole cause of lung cancer, accounts for approximately 54% of cases in the population. The higher the estimated prevalence of smoking, the higher the estimate of the attributable risk. The estimated $\widehat{V}[\log OR] = 0.04907$ then yields an estimated $V(\text{logit } PAR) = 0.0765$. This yields symmetric confidence limits on the $\text{logit}(PAR)$, which then yield asymmetric 95% confidence limits for the PAR of (0.412, 0.675).

5.2 MATCHING

We now turn to the case of a matched prospective study and a matched case-control study. Previously, a stratified analysis was employed to obtain an adjusted estimate

and a test of the risk difference between the treatment or exposure groups adjusting or controlling for the effect of an intervening covariate or confounding factor. As we shall see later, a stratified analysis is equivalent to using a regression model as the basis for the adjustment. An alternative approach to control for the effects of a covariate is matching. In this case, each member of a group (either E or \bar{E} in a prospective study, D or \bar{D} in a retrospective study) is matched to a member of the other group with respect to the values of one or more covariates Z .

5.2.1 Frequency Matching

One of the principal types of matching is frequency or within-class matching, in which elements or members of the comparison group are sampled within separate categories of a discrete covariate class such as gender (male/female) or decade of age (0–9 years, 10–19 years, etc.) so that the members of each group are matched within each category. For example, in a frequency-matched case-control study, the cases may be stratified jointly by gender and decade of age. Then within each category, such as males in their 40s, a separate sample of controls is selected from the control population in that category (males in their 40s). In this technique, any quantitative covariates such as age are grouped.

In a frequency-matched study, separate samples of cases and controls are obtained within covariate categories or strata. These samples need not be of equal size within or between strata. The objective is to have adequate numbers of subjects from each group within each stratum to provide an adequate overall comparison between groups. Thus, it is only necessary that a sufficient number of exposed and nonexposed cases and controls be sampled within each stratum so as to provide an assessment of the odds ratio or relative risk in aggregate over strata.

No single overall sample from each group is obtained at random from the population, such as an unrestricted sample of cases and an unrestricted sample of controls. Thus, the unadjusted marginal analysis has no simple corresponding likelihood, such as a simple product binomial. Rather, the underlying likelihood is a product binomial within strata, because separate samples are obtained within each stratum. Thus, the only appropriate analysis in this case is one that is stratified by the covariate categories used in the frequency matching. Of course, the analysis may be further stratified by additional unmatched covariates other than those used as the basis for the frequency-matched samples, such as by quintiles of body mass index in a study frequency matched by gender and decade of age.

Example 5.2 Frequency Matching

Korn (1984) describes a study of the association between a history of smoking (E vs. \bar{E}) and bladder cancer among males (D) versus an age frequency-matched sample

of noncancer control patients (\bar{D}). The study data are

Age Stratum	Cases		Controls	
	E	\bar{E}	E	\bar{E}
50–54	24	1	22	4
55–59	35	2	35	4
60–64	31	5	38	3
65–69	46	7	42	15
70–74	60	13	51	28
75–79	39	14	32	20
<i>Total</i>	235	42	220	74

(reproduced with permission). Again note that in a frequency-matched study such as this the number of matched controls sampled within a stratum need not equal the number of cases in that stratum, nor be proportional to the numbers of controls sampled in other strata.

The sampling procedure for this study was first to select all 277 cases (D) and to then stratify these cases by intervals of age. Within each age stratum, n_{2j} controls were then selected and the exposure status (E vs. \bar{E}) was determined for all cases and controls. This design, therefore, provides six separate independent samples (strata) with an independent 2×2 table within each stratum. Because the controls were sampled separately within strata, this design does not yield a single random sample of 277 cases and 294 controls. Thus, it is inappropriate to collapse the data into a single marginal 2×2 table. Rather, the proper analysis is a stratified analysis, such as a Mantel-Haenszel analysis of odds ratios.

The Mantel-Haenszel estimate of the common odds ratio is $\widehat{OR}_{MH} = 1.96$ with a Robins et al. estimated variance of $\widehat{V}[\log \widehat{OR}_{MH}] = 0.0483$. This yields a 95% C.I. for the odds ratio of (1.28, 3.02). The conditional Mantel-Haenszel test statistic is $X^2_{C(MH)} = 9.59$, with $p \leq 0.002$. Cochran's test of homogeneity is $X^2_{H(C)} = 4.38$ on 5 df , which is not statistically significant with $p \doteq 0.50$, indicating that a fixed effects model is appropriate.

5.2.2 Matched Pairs Design: Cross-Sectional or Prospective

The more common approach to matching is to construct matched pairs, or matched sets. In a cross-sectional or prospective study, matched sets are sampled, each consisting of an exposed (E) individual and one or more individuals who were not exposed (\bar{E}), where the members of the pair (or set) are uniquely matched with respect to a set of covariates. An example is the matched study where a control is selected from the population of nonexposed individuals who share the same covariates as each exposed individual, such as age and gender. Matched sets may also arise naturally, as when each subject serves as his or her own control and observations on each subject are obtained before (\bar{E}) and again after (E) exposure to a risk factor or the application of some treatment.

In some cases, the matching covariates are, in fact, unobservable quantities such as when the sample consists of a set of individuals with multiple measurements within-person. An example would be studies of matched organs within individuals (eyes, ears, kidneys, teeth, etc.), where one or more elements within each individual are exposed or treated with an experimental agent, and one or more are not. Another example is a study involving multiple family members, for example, among twins, where the family members have some genes in common. A similar example is when characteristics of littermates are studied. In this section, the methods for the analysis of matched data are described in terms of matched pairs of individuals, but they obviously apply to these other cases as well.

Consider a sample of N matched pairs with one exposed (E) and one nonexposed (\bar{E}) member, where the outcome of each member is designated as (D, \bar{D}) to denote developing the disease versus not, or having the positive versus negative outcome. In this setting it would be inappropriate to treat these as two independent groups in a 2×2 table such as

		E	\bar{E}	
		a	b	m_1
D	a			
	c		d	m_2
		N	N	$2N$

The principal reason that methods for two independent groups do not apply is that the observations within each pair are correlated so that we do not have a sample of $2N$ independent observations.

To see this, assume that the exposed and nonexposed members of each pair are matched with respect to some continuous covariate Z . We then assume that the probability of the outcome is some function of the covariate. Let Y_1 and Y_2 denote the responses for the exposed and nonexposed members, respectively. Then for the i th pair with shared covariate value z_i , let y_{ij} designate the outcome for the j th member, $y_{ij} = 1$ if D , 0 if \bar{D} , with probability $E(y_{ij}) = \pi_{ij}$ for $i = 1, \dots, N$ and $j = 1, 2$. To simplify, consider the case where the general null hypothesis is true such that the probabilities of the outcome are the same for each pair member conditional on the values z_i , that is, $\pi_{i1} = \pi_{i2}$. Denote these shared probabilities as $E(y_{ij}|z_i) = \pi(z_i)$ for the two members of the i th pair with covariate value z_i . Within the i th pair, the pair members are sampled independently so that the outcomes of the matched pairs y_{i1} and y_{i2} are *conditionally independent*. Thus, $E(y_{i1}|y_{i2}, z_i) = E(y_{i1}|z_i) = \pi(z_i)$ and $E(y_{i2}|y_{i1}, z_i) = E(y_{i2}|z_i) = \pi(z_i)$. The covariance of the matched measures over all sets is then provided by a generalization of (A.6)

$$Cov(Y_1, Y_2) = E_Z[Cov(Y_1, Y_2|Z)] + Cov[E(Y_1|Z), E(Y_2|Z)] \quad (5.20)$$

where $E_Z[Cov(Y_1, Y_2|Z)] = 0$ due to conditional independence and

$$Cov[E(Y_1|Z), E(Y_2|Z)] = Cov[\pi(z), \pi(z)] = Var[\pi(z)]. \quad (5.21)$$

From (A.6) it follows that

$$Cov(Y_1, Y_2) = E_Z[\pi(z)^2] - (E_Z[\pi(z)])^2 \neq 0, \quad (5.22)$$

so that Y_1 and Y_2 are correlated in the population of matched pairs (see Problem 5.3).

Thus, the proper analysis of matched pairs must allow for the within-pair correlation. In the case of a binary outcome this requires that we treat the pair as the sampling unit and then classify each pair according to the values of the outcomes of the two pair members. Thus, we have N pairs that are cross-classified as follows:

Frequencies				Probabilities				(5.23)	
E		\bar{E}		E		\bar{E}			
D	\bar{D}	D	\bar{D}	D	\bar{D}	D	\bar{D}		
E	\bar{D}	D	\bar{D}	π_{11}	π_{12}	$\pi_{1•}$			
D	\bar{D}	π_{21}	π_{22}	$\pi_{2•}$					
$n_{•1}$	$n_{•2}$	$\pi_{•1}$	$\pi_{•2}$				1		
	N								

Therefore, the sample frequencies are distributed as a multinomial (quadrinomial) where the likelihood of the sample is

$$P(e, f, g | N) = \frac{N!}{e! f! g! h!} \pi_{11}^e \pi_{12}^f \pi_{21}^g \pi_{22}^h. \quad (5.24)$$

As for the 2×2 table in Chapter 2, we use simple letters to refer to the frequencies in the table: $e = n_{11}$, $f = n_{12}$, and so on. The underlying probabilities of the four cells designate the probability of the outcomes for the exposed and nonexposed members of each pair. Thus, $\pi_{11} = P(D, D)$ for the exposed (E) and nonexposed (\bar{E}) pair members, respectively, $\pi_{12} = P(D, \bar{D})$, $\pi_{21} = P(\bar{D}, D)$, and $\pi_{22} = P(\bar{D}, \bar{D})$.

From the margins of the table it appears that $\pi_{1•}$ equals $P(D|E)$ and that $\pi_{•1}$ equals $P(D|\bar{E})$. However, $\pi_{1•}$ and $\pi_{•1}$ do not both refer to a conditional probability in the general population. Consider the case where matched pairs have been constructed on the basis of a continuous matching covariate Z . The covariate is distributed as $f_E(z) = P(z|E)$ within the exposed segment of the population, and as $f_{\bar{E}}(z) = P(z|\bar{E})$ within the nonexposed segment. If the covariate Z is, in fact, associated with the likelihood of exposure, then these distributions will differ between the exposed and nonexposed individuals, $f_E(z) \neq f_{\bar{E}}(z)$. This, in turn, implies the following developments.

Usually, we begin by selecting an exposed individual from its respective population. When the sample of exposed individuals can be viewed as a sample drawn at random from the exposed population, then $\pi_{1•} = P(D|E)$ irrespective of matching, since

$$\pi_{1•} = \int_z P(D|E, z) f_E(z) dz = P(D|E). \quad (5.25)$$

However, to construct a matched pair, first the covariate value z is identified for the exposed (E) member and then a control nonexposed individual \bar{E} is sampled at random conditional on z . That is, the control member is sampled from the subset of the nonexposed population who share the covariate value z . Thus, the distribution of the covariate Z among the sampled nonexposed individuals equals the distribution among the exposed individuals, so that

$$\pi_{•1} = \int_z P(D|\bar{E}, z) f_E(z) dz = P_m(D|\bar{E}), \quad (5.26)$$

where $P_m(D|\bar{E})$ is used to denote the conditional probability under matching. However, since the distribution of the matching covariate differs among those exposed and those not exposed, $f_E(z) \neq f_{\bar{E}}(z)$, then this conditional probability $P_m(D|\bar{E})$ is not the same as the conditional probability in the entire population of nonexposed individuals

$$\pi_{\bullet 1} \neq \int_z P(D|\bar{E}, z) f_{\bar{E}}(z) dz = P(D|\bar{E}). \quad (5.27)$$

Therefore, measures of association such as the relative risk or odds ratio based on the marginal probabilities from the matched or paired 2×2 table do not have an unconditional population interpretation. Rather, any such measures must be interpreted conditionally on the matching on other covariates.

5.3 TESTS OF ASSOCIATION FOR MATCHED PAIRS

Before considering measures of association for matched data, we first consider tests of the hypothesis of no association. For the paired or matched 2×2 table there are two equivalent hypotheses of interest. The first is the hypothesis of *marginal homogeneity* under matching,

$$H_0: P(D|E) = P_m(D|\bar{E}) \text{ or } \pi_{1\bullet} = \pi_{\bullet 1}. \quad (5.28)$$

Since this hypothesis is stated in terms of the marginal probabilities from an underlying quadrinomial in (5.23), then this hypothesis implies and is implied by the hypothesis of *symmetry*,

$$H_0: \pi_{12} = \pi_{21} \quad (5.29)$$

with respect to the discordant pairs.

5.3.1 Exact Test

The likelihood of the 2×2 table for matched pairs is a quadrinomial. However, we wish to conduct a test of the equality of two of the four probabilities in this likelihood. To do so we must account for or eliminate the other nuisance parameters (π_{11}, π_{22}) . This is readily done through the principle of conditioning, as illustrated in Section 2.4.2, through the derivation of the conditional hypergeometric distribution. This approach requires that we identify a conditional distribution that only involves the parameters of interest (π_{12}, π_{21}) .

It is readily shown in Problem 5.7 that the conditional distribution of f given the number of discordant pairs $M = f + g$ is a simple binomial distribution

$$P(f | M) = \frac{M!}{f!g!} \left(\frac{\pi_{12}}{\pi_d} \right)^f \left(\frac{\pi_{21}}{\pi_d} \right)^g = B\left(f; M, \frac{\pi_{12}}{\pi_d}\right), \quad (5.30)$$

where $\pi_d = \pi_{12} + \pi_{21}$ is the probability of a discordant pair in which one member of the pair has a positive outcome, the other a negative outcome.

Therefore, under the null hypothesis of symmetry $H_0: \pi_{12} = \pi_{21}$, it follows that

$$P(f | M) = B(f; M, 1/2). \quad (5.31)$$

Therefore, an exact two-sided p -value is obtained as

$$p = 2 \sum_{j=0}^{\min(f,g)} B(j; M, 1/2), \quad (5.32)$$

which is a simple binomial test.

Example 5.3 Small Sample

For example consider a paired 2×2 table with discordant entries

-	2
8	-

To test the hypothesis of symmetry, the total sample size and the numbers of concordant pairs are irrelevant. Conditioning on the number of discordant pairs, $M = 10$, then the two-sided exact p -value is

$$\begin{aligned} p &= 2[B(0; 10, 1/2) + B(1; 10, 1/2) + B(2; 10, 1/2)] \\ &= 2[0.001 + 0.0097 + 0.0440] = 0.1094. \end{aligned}$$

Note that this quantity can also be computed by hand simply as

$$p = 2(0.5)^{10} \left[\binom{10}{0} + \binom{10}{1} + \binom{10}{2} \right] = 2(0.5)^{10} [1 + 10 + 45] = 0.1094.$$

5.3.2 McNemar's Large Sample Test

A large sample test can be derived either from the normal approximation to the multinomial, or from the normal approximation to the conditional binomial distribution. Here we use the former while the derivation using the latter is left to Problem 5.5.

Let the proportion within the ij th cell be $p_{ij} = n_{ij}/N$, where $E(p_{ij}) = \pi_{ij}$, $i = 1, 2$; $j = 1, 2$. From the normal approximation to the multinomial, the vector of cell proportions $\mathbf{p} = (p_{11} \ p_{12} \ p_{21} \ p_{22})^T$ is asymptotically normally distributed with expectation $\boldsymbol{\pi} = (\pi_{11} \ \pi_{12} \ \pi_{21} \ \pi_{22})^T$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\pi})/N$, as shown in Example A.2, where the subscripts (11, 12, 21, 22) are equivalent to (1, 2, 3, 4). Because the proportions sum to 1.0, the distribution is degenerate, meaning that the covariance matrix is singular with rank 3. However, this distribution can still be used as the basis for the evaluation of the distribution of contrasts among the proportions.

Clearly, to test the hypothesis of symmetry, $H_0: \pi_{12} = \pi_{21}$, we wish to use the test statistic based on the difference in the discordant proportions $p_{12} - p_{21}$ of the form

$$Z = \frac{p_{12} - p_{21}}{\sqrt{\hat{V}(p_{12} - p_{21}|H_0)}} \quad (5.33)$$

with the variance of the difference evaluated under the null hypothesis. Using a linear contrast in p_{12} and p_{21} , it is readily shown (Problem 5.6) that

$$V(p_{12} - p_{21}) = \frac{(\pi_{12} + \pi_{21}) - (\pi_{12} - \pi_{21})^2}{N}. \quad (5.34)$$

Under H_0 : $\pi_{12} = \pi_{21} = \pi$, where $\pi = \pi_d/2$,

$$V(p_{12} - p_{21}|H_0) = \frac{\pi_{12} + \pi_{21}}{N} = \frac{\pi_d}{N} \hat{=} \frac{p_{12} + p_{21}}{N}. \quad (5.35)$$

Therefore,

$$Z_M = \frac{p_{12} - p_{21}}{\sqrt{(p_{12} + p_{21})/N}} = \frac{f - g}{\sqrt{f + g}} \quad (5.36)$$

is asymptotically normally distributed under H_0 . This statistic provides either a one- or a two-sided test of H_0 . For a two-sided test one can use the equivalent chi-square statistic

$$X_M^2 = \frac{(f - g)^2}{f + g}, \quad (5.37)$$

which is asymptotically distributed as chi-square on 1 *df*. This is *McNemar's test* (McNemar, 1947).

Example 5.4 Large Sample

Consider the paired 2×2 table with the discordant frequencies

-	80
55	-

where $M = 135$, $f = 80$, and $g = 55$. McNemar's test of the hypothesis of symmetry yields $X_M^2 = (25)^2/135 = 4.63$ with $p \leq 0.032$.

5.3.3 SAS PROC FREQ

The SAS PROC FREQ provides for the computation of the exact and large sample McNemar test for matched or paired 2×2 tables. Table 5.1 presents the SAS statements to conduct an analysis of the data in Example 5.4. While only the discordant observations are required to compute the exact and large sample tests, SAS requires that the complete table be provided. Thus, the rows are labeled using the variable *Emember*, that designates the response of the exposed member of each pair and the variable *Nmember*, that designates that for the nonexposed member. Unfortunately, all of the other various tests and measures for unmatched 2×2 tables are also provided in the SAS output, that can be highly misleading. Further, the measures of association for matched pairs are not presented.

Table 5.1 SAS PROC FREQ analysis of Example 5.4.

```

data one; input e f g h;
**** Note that the values of e and h are irrelevant;
cards;
10 80 55 10
;
title1 'Chapter 5, Example 5.4';
data two; set one;
Emember=1; Nmember=1; x=e; output;
Emember=1; Nmember=2; x=f; output;
Emember=2; Nmember=1; x=g; output;
Emember=2; Nmember=2; x=h; output;
proc freq; tables Emember* Nmember / all nopercent nocol agree;
  weight x; exact mcnem;
title2 'SAS PROC FREQ Analysis of Matched 2 x 2 tables';
run;

```

5.4 MEASURES OF ASSOCIATION FOR MATCHED PAIRS

We now consider the definition and estimation of measures of association for matched pairs, starting with the odds ratio.

5.4.1 Conditional Odds Ratio

In the general unselected population, the population odds ratio is defined as in (5.11), where the conditional probabilities, such as $P(D|E)$, are defined for the general population. However, the marginal odds ratio under matching is

$$OR_m = \frac{\pi_{1\bullet}/\pi_{2\bullet}}{\pi_{\bullet 1}/\pi_{\bullet 2}} = \frac{P(D|E)/P(\bar{D}|E)}{P_m(D|\bar{E})/P_m(\bar{D}|\bar{E})} \neq OR. \quad (5.38)$$

Therefore, the observed marginal odds ratio does not provide an estimate of the population odds ratio. Thus, this is called the *population averaged odds ratio* because both $\pi_{1\bullet}$ and $\pi_{\bullet 1}$ are the average risks or the expectations of $P(D|E, z)$ and $P(D|\bar{E}, z)$, respectively, with respect to the distribution of the matching covariate in the exposed population, $f_E(z)$.

Rather than attempting to describe the odds ratio unconditionally, let's do so conditionally with respect to the distribution of the covariate among the exposed. Conditionally, for a specific value of the covariate z , the conditional odds ratio is

$$OR_z = \frac{P(D|E, z)P(\bar{D}|\bar{E}, z)}{P(D|\bar{E}, z)P(\bar{D}|E, z)}. \quad (5.39)$$

Given z , the matched E and \bar{E} pair members are sampled independently from their respective populations, and thus the pair members are conditionally independent. Therefore, for each pair with covariate value z , a 2×2 table can be constructed such as that in (5.23) with cell probabilities $\{\pi_{ij|z}\}$, row marginal probabilities $\{\pi_{i*|z}\}$, and column marginal probabilities $\{\pi_{*j|z}\}$ for $i = 1, 2$; $j = 1, 2$. Since the pairs are conditionally independent, then within each such table

$$\pi_{12|z} = \pi_{1*|z} \pi_{*2|z} = P(D|E, z) P(\bar{D}|\bar{E}, z), \quad (5.40)$$

and

$$\pi_{21|z} = \pi_{2*|z} \pi_{*1|z} = P(D|\bar{E}, z) P(\bar{D}|E, z), \quad (5.41)$$

so that the conditional marginal odds ratio is

$$OR_z = \frac{\pi_{1*|z} \pi_{*2|z}}{\pi_{2*|z} \pi_{*1|z}} = \frac{\pi_{12|z}}{\pi_{21|z}}. \quad (5.42)$$

Now assume that although $P(D|E, z)$ and $P(D|\bar{E}, z)$ may vary with z , there is a constant conditional marginal odds ratio for all values of the matching covariate z (Ejigou and McHugh, 1977, among others) such that

$$OR_z = OR_C \quad \forall z. \quad (5.43)$$

This implies that the population-averaged discordant probabilities π_{12} and π_{21} , each averaged with respect to the distribution of the covariate in the exposed population, then satisfy

$$\pi_{21} = \int_z \pi_{21|z} f_E(z) dz = \int_z \pi_{2*|z} \pi_{*1|z} f_E(z) dz, \quad (5.44)$$

and from (5.42),

$$\pi_{12} = \int_z OR_z \pi_{21|z} f_E(z) dz = OR_C \int_z \pi_{2*|z} \pi_{*1|z} f_E(z) dz = OR_C \pi_{21}. \quad (5.45)$$

Thus, the conditional odds ratio is

$$OR_C = \frac{\pi_{12}}{\pi_{21}} \doteq \frac{p_{12}}{p_{21}} = \frac{f}{g}. \quad (5.46)$$

In Chapter 6 we show that this quantity also arises from Cox's (1958b) conditional logit model for pair-matched data.

5.4.2 Confidence Limits for the Odds Ratio

5.4.2.1 Exact Limits Exact confidence limits for the conditional odds ratio can be obtained from the conditional binomial distribution presented in (5.30). Thus,

$f \sim B(f; M, \pi_f)$, where $\pi_f = \pi_{12}/\pi_d$ and $g \sim B(g; M, \pi_g)$, where $\pi_g = \pi_{21}/\pi_d$. Since $OR_c = \pi_{12}/\pi_{21}$, then

$$OR_c = \frac{\pi_f}{1 - \pi_f} = \frac{1 - \pi_g}{\pi_g}. \quad (5.47)$$

Let $a = \min(f, g)$. Then the upper confidence limit at level $1 - \alpha$ for $\pi_a = \min(\pi_f, \pi_g)$ is that value of $\pi_{a(u)}$ such that

$$\alpha/2 = \sum_{x=0}^a \binom{M}{x} \pi_{a(u)}^x (1 - \pi_{a(u)})^{M-x}. \quad (5.48)$$

Likewise, the lower limit $\pi_{a(\ell)}$ satisfies

$$\alpha/2 = \sum_{x=a}^M \binom{M}{x} \pi_{a(\ell)}^x (1 - \pi_{a(\ell)})^{M-x}. \quad (5.49)$$

Example 5.5 Small Sample

Consider the above example with $f = 2$ and $M = 10$. Then using StatXact the 95% confidence limits (0.02521, 0.5561) are obtained as the solution to the equations

$$\sum_{x=2}^{10} B(x; 10, \pi_{f(\ell)}) = 0.025, \quad (5.50)$$

and

$$\sum_{x=0}^2 B(x; 10, \pi_{f(u)}) = 0.025. \quad (5.51)$$

5.4.2.2 Large Sample Limits To obtain large sample confidence limits for the odds ratio, as for independent (unmatched) observations, it is preferable to use $\theta = \log OR_c$, which can be estimated consistently as $\hat{\theta} = \log(f/g)$. From the normal approximation to the multinomial, using the δ -method it can be shown that

$$V(\hat{\theta}) = \frac{1}{N} \left(\frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} \right) = \frac{\pi_d}{N\pi_{12}\pi_{21}}. \quad (5.52)$$

Because the cell proportions provide consistent estimates of the cell probabilities, the variance can be estimated consistently as

$$\hat{V}(\hat{\theta}) = \frac{1}{N} \left(\frac{N}{f} + \frac{N}{g} \right) = \left(\frac{1}{f} + \frac{1}{g} \right) = \frac{M}{fg}, \quad (5.53)$$

which only involves the discordant frequencies f and g , where $M = f + g$. This large sample estimated variance can then be used to construct asymmetric confidence limits on the conditional odds ratio OR_c .

5.4.3 Conditional Large Sample Test and Confidence Limits

Alternatively, McNemar's test and the asymmetric confidence limits for the odds ratio can be obtained from the large sample normal approximation to the conditional distribution of f given M . Under H_1 in (5.30) we showed that $f \sim B(f; M, \pi_{12}/\pi_d)$, whereas under H_0 , $f \sim B(f; M, 1/2)$. Therefore, under H_0 the large sample approximation to the binomial in terms of the proportion $p_f = f/M$ yields

$$H_0: p_f \stackrel{d}{\approx} \mathcal{N}\left(\frac{1}{2}, \frac{1}{4M}\right), \quad (5.54)$$

and the square of the large sample Z -test equals McNemar's test in (5.37) (see Problem 5.7).

Under H_1 the large sample approximation yields

$$H_1: p_f \stackrel{d}{\approx} \mathcal{N}\left[\pi_f, \frac{\pi_f(1 - \pi_f)}{M}\right], \quad (5.55)$$

where $\pi_f = \pi_{12}/\pi_d$. This then provides the basis for large sample confidence limits on π_f . Although one can compute the usual symmetric confidence limits for this probability, the asymmetric limits based on the logit transformation as presented in (2.15–2.16) are preferred. Note, however, that $\text{logit}(\pi_f) = \log(\pi_{12}/\pi_{21}) = \log OR_c$. Thus, in Problem 5.7 it is also shown that the confidence limits for the logit of π_f equal those for the conditional odds based on $\widehat{V}(\widehat{\theta})$ presented in (5.53).

Example 5.6 Large Sample

For the large sample data set in Example 5.4, the estimate of the conditional odds ratio is $\widehat{OR}_C = 80/55 = 1.45$ and the estimated log odds ratio is $\widehat{\theta} = 0.375$ with $S.E.(\widehat{\theta}) = \sqrt{135/(80 \times 55)} = 0.1752$. Therefore, the asymmetric 95% confidence limits for OR_C are $\exp[0.375 \pm (1.96 \times 0.1752)] = (1.03, 2.05)$.

5.4.4 Mantel-Haenszel Analysis

Another way to approach the problem of matched pairs was proposed by Mantel and Haenszel (1959). Since each pair has two members, each sampled independently given the value of the matching covariate Z , then the members within each pair are conditionally independent. Thus, they suggested that the sample of matched pairs be considered to consist of N independent samples (strata) consisting of one member in each exposure group (E , \bar{E}) in a cross-sectional or prospective study. These N samples provide N independent 2×2 tables each comprising one matched pair. The i th pair then provides an unmatched 2×2 table of the form

	E	\bar{E}	
D	a_i	b_i	m_{1i}
\bar{D}	c_i	d_i	m_{2i}
	1	1	2

(5.56)

Table 5.2 Mantel-Haenszel stratified analysis of pair-matched data.

#:	Tables of Each Type																			
	e	f	g	h																
Table:	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>1</td><td>1</td></tr><tr><td>0</td><td>0</td></tr></table>	1	1	0	0	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>1</td><td>0</td></tr><tr><td>0</td><td>1</td></tr></table>	1	0	0	1	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>0</td><td>1</td></tr><tr><td>1</td><td>0</td></tr></table>	0	1	1	0	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td></tr></table>	0	0	1	1
1	1																			
0	0																			
1	0																			
0	1																			
0	1																			
1	0																			
0	0																			
1	1																			
a_i :	1	1	0	0																
$E(a_i)$:	1	1/2	1/2	0																
$V_c(a_i)$:	0	1/4	1/4	0																

where $m_{1i} = 0, 1$, or 2 . In a case-control study similar tables are constructed from each pair consisting of one member from each disease or outcome group (D, \bar{D}). A stratified Mantel-Haenszel analysis can then be performed over these N tables.

However, it is readily shown (Problem 5.9) that there are only four possible types of 2×2 tables, each with values a_i , $E(a_i)$, and $V_c(a_i)$, as shown in Table 5.2. It can then be shown (Problem 5.9) that the summary Mantel-Haenszel test statistic X_{MH}^2 equals McNemar's test statistic X_M^2 in (5.37) and that the Mantel-Haenszel estimated summary odds ratio in (4.14) is the conditional odds ratio: $\widehat{OR}_{MH} = f/g = \widehat{OR}_C$. Further, it is also readily shown that the Robins et al. (1986) estimate of the variance of the log odds ratio in (4.28) equals the estimate of the large sample variance of the log conditional odds ratio presented in (5.53). Thus, a Mantel-Haenszel analysis of odds ratios from the N pair-matched tables provides results that are identical to those described previously for the conditional odds ratio.

5.4.5 Relative Risk for Matched Pairs

Like the odds ratio, the relative risk in the general population does not equal the marginal relative risk under matching. When the nonexposed controls are matched to the exposed members of each pair, as shown in Section 5.2.2, the *population-averaged relative risk* is

$$RR_A = \frac{\pi_{1\bullet}}{\pi_{\bullet 1}} = \frac{P(D|E)}{P_m(D|\bar{E})} \neq \frac{P(D|E)}{P(D|\bar{E})} = RR, \quad (5.57)$$

where RR_A is estimated as

$$\widehat{RR}_A = \frac{p_{1\bullet}}{p_{\bullet 1}} = \frac{p_{11} + p_{12}}{p_{11} + p_{21}} = \frac{e + f}{e + g} = \frac{n_{1\bullet}}{n_{\bullet 1}} \quad (5.58)$$

using the entries in the aggregate paired 2×2 table. In Problem 5.9, we also show that the Mantel-Haenszel estimate of the relative risk in (4.17) obtained from the N conditionally independent pairs equals this quantity.

Exact confidence limits are not readily obtained because the parameter involves the probabilities $(\pi_{11}, \pi_{12}, \pi_{21})$. Large sample asymmetric confidence limits can be obtained using the large sample variance of $\log(\widehat{RR}_A)$. In Problem 5.6.6, using the δ -method, it is shown that

$$V \left[\log(\widehat{RR}_A) \right] = \frac{\pi_{12} + \pi_{21}}{N\pi_{1\bullet}\pi_{\bullet 1}} = \frac{\pi_d}{N\pi_{1\bullet}\pi_{\bullet 1}}, \quad (5.59)$$

that can be estimated consistently as

$$\widehat{V} \left[\log(\widehat{RR}_A) \right] = \frac{p_{12} + p_{21}}{Np_{1\bullet}p_{\bullet 1}} = \frac{M}{n_{1\bullet}n_{\bullet 1}}. \quad (5.60)$$

5.4.6 Attributable Risk for Matched Pairs

Although the population attributable risk is a function of the relative risk, it is not a function of the population-averaged relative risk as estimated from a matched prospective study. Thus, the *PAR* is rarely used to describe the results of a matched cross-sectional or prospective study.

Example 5.7 *Pregnancy and Retinopathy Progression*

So far our discussion of matched pairs has concerned separate individuals that are matched to construct each pair. However, matched pairs may also arise from repeated measures on the same subject, among other situations. An example is provided by the study of the effects of pregnancy on progression of retinopathy (diabetic retinal disease) among women who became pregnant during the Diabetes Control and Complications Trial (DCCT, 2000). The retinopathy status of each woman was assessed before pregnancy and then again during pregnancy. Retinopathy was assessed on a 13-step ordinal scale of severity. Thus, rather than a 2×2 table to describe the retinopathy level before pregnancy and that during pregnancy, a 13×13 table is required. However, the study wished only to assess whether pregnancy exacerbated the level of retinopathy represented by *any* worsening versus *any* improvement. In this case, worsening refers to any cell above the diagonal of this 13×13 table, and improvement by any cell below.

For example, if retinopathy were assessed with four steps or ordinal categories, the data would be summarized in a 4×4 table such as

		Level During Pregnancy			
		1	2	3	4
Level Before Pregnancy	1	=	w	w	w
	2	b	=	w	w
	3	b	b	=	w
	4	b	b	b	=

All entries along the diagonal refer to individuals for whom the level before pregnancy equals that during pregnancy, where pregnancy is equivalent to the risk factor exposure. Then the total number of individuals worse, say W , is the sum of the

entries in all cells above the diagonal of equality, and the total number improved or better, say B , is the sum of the entries in all cells below the diagonal. These quantities are equivalent to the discordant frequencies used previously such that $f = W$ and $g = B$. Then, from multinomial distribution for the 16 cells of the table, or by conditioning on the total number of discordant observations, $M = W + B$, it can be shown that all of the above results apply.

Note that one can do likewise for any symmetric partitioning of the entries in the paired two-way table. Thus, if one is only interested in worsening by two or more steps above the diagonal, only the entries in the three extreme cells in either corner of the table would be used.

The effects of pregnancy on the level of retinopathy in the DCCT are then presented in the following table:

		During	
		-	$W = 31$
Before	-	$W = 31$	
	$B = 14$	-	

Among the 77 women who became pregnant, 31 (40%) had a level of retinopathy that was worse during pregnancy than before pregnancy, compared to 14 (18%) whose level of retinopathy improved. Under the null hypothesis of no effect of pregnancy, the numbers worse and better are expected to be equal. The conditional odds ratio is $\widehat{OR}_C = 2.2$ with asymmetric 95% C.I. (1.18, 4.16). McNemar's test yields a value $X_M^2 = 6.42$ with $p \leq 0.012$.

5.5 PAIR-MATCHED RETROSPECTIVE STUDY

One of the most common applications of matched pairs is in a retrospective case-control study. Here the sampling involves the selection of a case with the disease (D), the determination of the covariate values for that subject (z), the selection of a matched nondiseased control (\bar{D}) with the same covariate values, and then the determination of whether the case and control had previously been exposed (E) or not (\bar{E}) to the risk factor under study. In this setting the observations are summarized in a 2×2 table as

		Frequencies				Probabilities			
		D				\bar{D}			
		E	\bar{E}			E	\bar{E}		
D	E	e	f	$n_{1\bullet}$		D	E	ϕ_{11}	ϕ_{12}
	\bar{E}	g	h	$n_{2\bullet}$			\bar{E}	ϕ_{21}	ϕ_{22}
		$n_{\bullet 1}$	$n_{\bullet 2}$	N				$\phi_{\bullet 1}$	$\phi_{\bullet 2}$
								1	

(5.61)

As was the case for the prospective study, the marginal probability $\phi_{1\bullet}$ under matching can be expressed as $P(E|D) = \int_z P(E|D, z) f_D(z) dz$ with respect to the

distribution of Z among the cases. However, since the controls are selected to have covariate values that match those of the cases, then

$$\begin{aligned}\phi_{\bullet 1} &= \int_z P(E|\bar{D}, z) f_D(z) dz = P_m(E|\bar{D}), \\ &\neq \int_z P(E|\bar{D}, z) f_{\bar{D}}(z) dz = P(E|\bar{D})\end{aligned}\quad (5.62)$$

analogous to (5.26) for the matched prospective study. Therefore, the marginal retrospective odds ratio under matching does not equal the retrospective odds ratio in the general population of cases and controls, and thus does not equal the prospective odds ratio.

As for the prospective study the null hypothesis of marginal homogeneity under matching $H_0: \phi_{1\bullet} = \phi_{\bullet 1}$ implies the hypothesis of symmetry $H_0: \phi_{12} = \phi_{21}$. Further, the hypothesis of symmetry of the discordant retrospective probabilities implies the null hypothesis of symmetry of the prospective discordant probabilities (Problem 5.10). Thus, the conditional binomial exact test and the large sample McNemar test are again employed to test the hypothesis of no association between exposure and disease.

5.5.1 Conditional Odds Ratio

As in Section 5.4.1, conditional on the value z of the matching covariate, the pair members are sampled independently such that the probability of each type of discordant pair is

$$\phi_{12|z} = \phi_{1\bullet|z} \phi_{\bullet 2|z} = P(E|D, z) P(\bar{E}|\bar{D}, z) \quad (5.63)$$

and

$$\phi_{21|z} = \phi_{2\bullet|z} \phi_{\bullet 1|z} = P(\bar{E}|D, z) P(E|\bar{D}, z). \quad (5.64)$$

Thus, the conditional marginal retrospective odds ratio is

$$OR_z = \frac{\phi_{1\bullet|z} \phi_{\bullet 2|z}}{\phi_{2\bullet|z} \phi_{\bullet 1|z}} = \frac{\phi_{12|z}}{\phi_{21|z}}. \quad (5.65)$$

For the unmatched retrospective study we showed in (5.11) that the retrospective odds ratio equals the prospective odds ratio. For the matched retrospective study, it can likewise be shown that the conditional retrospective odds ratio

$$OR_z = \frac{P(E|D, z) / P(\bar{E}|D, z)}{P(E|\bar{D}, z) / P(\bar{E}|\bar{D}, z)} \quad (5.66)$$

equals the conditional prospective odds ratio

$$OR_z = \frac{P(D|E, z) / P(\bar{D}|E, z)}{P(D|\bar{E}, z) / P(\bar{D}|\bar{E}, z)}, \quad (5.67)$$

using the same developments as for the matched prospective study (Problem 5.10).

Also, the assumption of a common conditional retrospective odds ratio implies a common conditional prospective odds ratio,

$$OR_z = \underset{retro}{OR_C} = \underset{prosp}{OR_C} \quad \forall z, \quad (5.68)$$

which is estimated consistently as $\widehat{OR}_C = f/g$ with estimated large sample variance $\widehat{V}[\log \widehat{OR}_C] = M/(fg)$. This then provides for the computation of large sample confidence limits.

5.5.2 Relative Risks from Matched Retrospective Studies

For the unmatched retrospective study, Cornfield (1951) showed that the relative risk can be approximated by the odds ratio under the rare disease assumption. This result also applies to the matched retrospective study under matching. In this case it can be shown that the relative risk for a fixed value z of the matching covariate is

$$\begin{aligned} RR_z &= \frac{P(D|E, z)}{P(D|\bar{E}, z)} = \frac{P(E|D, z)P(\bar{E}|z)}{P(\bar{E}|D, z)P(E|z)} \\ &= \frac{P(E|D, z)P(\bar{E}|D, z)P(D|z) + \phi_{12|z}[1 - P(D|z)]}{P(E|D, z)P(\bar{E}|D, z)P(D|z) + \phi_{21|z}[1 - P(D|z)]}, \end{aligned} \quad (5.69)$$

where $\phi_{12|z}$ and $\phi_{21|z}$ are the discordant probabilities conditional on the value z and $P(D|z) = \delta_z$ is the conditional probability of the disease in the population given z . Again assuming a constant relative risk for all values of Z , then under the rare disease assumption ($\delta_z \downarrow 0$),

$$RR_{z, \delta_z \downarrow 0} \doteq \frac{\phi_{12|z}}{\phi_{21|z}} = OR_C, \quad (5.70)$$

so that the conditional odds ratio provides an approximation to the relative risk.

Example 5.8 Estrogen Use and Endometrial Cancer

Breslow and Day (1980) present data from Mack et al. (1976) for a matched case-control study of the risk of endometrial cancer among women who resided in a retirement community in Los Angeles. The data set is described in Problem 7.18. Each of the 63 cases was matched to four controls within one year of age of the case, who had the same marital status and who entered the community at about the same time. Here only the first matched control is employed to construct 63 matched pairs. The risk factor exposure of interest is a history of use of conjugated estrogens. The resulting table of frequencies is

		\overline{D}		51
		E	\overline{E}	
D	E	18	33	51
	\overline{E}	6	6	
		24	39	63

for which McNemar's test is $X_M^2 = 18.69$ with $p < 0.0001$. The estimate of the conditional odds ratio is $\widehat{OR}_C = 5.5$ with asymmetric 95% confidence limits (2.30, 13.13). This odds ratio provides an approximate relative risk under the rare disease assumption.

5.6 POWER FUNCTION OF McNEMAR'S TEST

5.6.1 Unconditional Power Function

Various authors have proposed methods for evaluation of the power of the McNemar test, some unconditionally assuming only a sample of N matched pairs, others conditionally assuming that M discordant pairs are observed. Among the unconditional power function derivations, Lachin (1992b) shows that the procedure suggested by Connor (1987) and Connett et al. (1987) based on the underlying multinomial is the most accurate. This approach will be described for a prospectively matched study (in terms of the π_{ij}), but the same equations also apply to the matched retrospective study (in terms of the ϕ_{ij}).

The test statistic is $T = p_{12} - p_{21}$. The null hypothesis of symmetry specifies that $H_0: \pi_{12} = \pi_{21} = \pi_d/2$, where π_d is the probability of a discordant pair. Then, from (5.33) and (5.35), $T \stackrel{d}{\approx} \mathcal{N}(\mu_0, \sigma_0^2)$ with $\mu_0 = 0$ and $\sigma_0^2 = \pi_d/N$. Under the alternative hypothesis $H_1: \pi_{12} \neq \pi_{21}$, from (5.34), $T \stackrel{d}{\approx} \mathcal{N}(\mu_1, \sigma_1^2)$ with $\mu_1 = \pi_{12} - \pi_{21}$ and

$$\sigma_1^2 = \frac{(\pi_{12} + \pi_{21}) - (\pi_{12} - \pi_{21})^2}{N} = \frac{\pi_d - (\pi_{12} - \pi_{21})^2}{N}. \quad (5.71)$$

Using the general equation relating sample size to power in (3.18), it follows that

$$\sqrt{N} |\pi_{12} - \pi_{21}| = Z_{1-\alpha} \sqrt{\pi_d} + Z_{1-\beta} \sqrt{\pi_d - (\pi_{12} - \pi_{21})^2}, \quad (5.72)$$

from which one can determine N or solve for the standardized deviate corresponding to the level of power $Z_{1-\beta}$ (see Problem 5.11).

To use this expression, one only need specify π_{21} and $OR = \pi_{12}/\pi_{21}$ so that $\pi_{12} = \pi_{21}OR$. The other entries in the 2×2 table are irrelevant.

Example 5.9 *Unconditional Sample Size*

Assume that $\pi_{21} = 0.125$ and that we wish to detect an odds ratio of $OR = 2$, which, in turn, implies that $\pi_{12} = 0.25$, $\pi_d = 0.375$, and $\mu_1 = 0.125$. For $\alpha = 0.05$ (two-sided) the N required to provide $\beta = 0.10$ is provided by

$$N = \left[\frac{1.96\sqrt{0.375} + 1.282\sqrt{(0.375) - (0.125)^2}}{0.125} \right]^2 = 248.$$

5.6.2 Conditional Power Function

The unconditional power function above is the preferred approach for determining the sample size a priori. The actual power, however, depends on the observed number of discordant pairs, $M = f + g$. In the above example, given the specified values of the discordant probabilities, the expected number of discordant pairs is $E(M) = N\pi_d = (248)(0.375) = 93$. It is the latter quantity that determines the power of the test, as is demonstrated from the expressions for the conditional power function.

Miettinen (1968) described the conditional power function given M in terms of the statistic $(f - g)$. However, it is simpler to do this in terms of the conditional distribution of $p_f = f/M$ described in Section 5.4.3. Under H_0 in (5.54), $p_f \stackrel{d}{\approx} \mathcal{N}(\mu_{0c}, \sigma_{0c}^2)$ with $\mu_{0c} = 1/2$ and $\sigma_{0c}^2 = 1/(4M)$. Conversely, under the alternative H_1 in (5.55), $p_f \stackrel{d}{\approx} \mathcal{N}(\mu_{1c}, \sigma_{1c}^2)$ with $\mu_{1c} = \pi_{12}/\pi_d$ and

$$\sigma_{1c}^2 = \frac{\pi_{12}\pi_{21}}{M(\pi_d)^2} \quad (5.73)$$

(see Problem 5.11). These expressions can also be presented in terms of the conditional odds ratio, say $\varphi_c = OR_c$ (see Problem 5.11.3). Since $\varphi_c = \pi_{12}/\pi_{21}$, then

$$\mu_{1c} = \frac{\pi_{12}}{\pi_d} = \frac{\pi_{12}}{\pi_{12} + \pi_{21}} = \frac{\varphi_c}{1 + \varphi_c}, \quad (5.74)$$

and

$$\sigma_{1c}^2 = \frac{\pi_{12}\pi_{21}}{M(\pi_d)^2} = \frac{\varphi_c}{M(1 + \varphi_c)^2}. \quad (5.75)$$

Substituting into the general expression (3.20) for $Z_{1-\beta}$, the conditional power function is provided by

$$Z_{1-\beta} = \frac{\sqrt{M} |\pi_{12} - \pi_{21}| - Z_{1-\alpha} \pi_d}{2\sqrt{\pi_{12}\pi_{21}}}. \quad (5.76)$$

When specified in terms of M and φ_c , we obtain

$$Z_{1-\beta} = \frac{\sqrt{M} |\varphi_c - 1| - Z_{1-\alpha} (\varphi_c + 1)}{2\sqrt{\varphi_c}}, \quad (5.77)$$

which provides the level of conditional power $1 - \beta$ to detect an odds ratio of φ_c with M discordant pairs.

The conditional power function should not be used to determine the sample size a priori unless the study design calls for continued sampling until the fixed number M of discordant pairs is obtained. Otherwise, there is no guarantee that the required M will be observed. Therefore, the unconditional expressions based on (5.72) should be used to determine the total sample size N to be sampled a priori. This is the value N that on average will provide the required number of discordant pairs for the specified values of π_{21} and φ_c .

Example 5.10 Conditional Power

For Example 5.9, where we wished to detect an odds ratio $\varphi_c = 2$, if the actual number of discordant pairs is $M = 80$ rather than the desired 93, then the standardized deviate is

$$Z_{1-\beta} = \frac{\sqrt{80}|2 - 1| - (1.96)(3)}{2\sqrt{2}} = 1.083, \quad (5.78)$$

for which the corresponding level of power is $1 - \beta \doteq 0.86$.

5.6.3 Other Approaches

Schlesselman (1982) describes an unconditional calculation based on an extension of the above expression for the conditional power function in a matched case-control study with probabilities $\{\phi_{ij}\}$ in the four cells of the table, where the $\{\phi_{ij}\}$ in the retrospective study are the counterparts to the $\{\pi_{ij}\}$ in the prospective study. His approach is appealing because the problem is parameterized in terms of the *marginal* probabilities, that are often known, rather than the discordant probabilities, that are usually unknown. Thus, his approach is based on specification of the odds ratio to be detected and the marginal population probabilities of exposure among the controls, $\phi_{1\bullet} = P(E|\bar{D})$. However, Schlesselman's procedure assumes that the responses of the pair members are independent, or, for example, that $\phi_{11} = \phi_{1\bullet}\phi_{\bullet 1}$, which only occurs when matching is ineffective.

Various authors have provided corrections to Schlesselman's calculation based on the specification of the correlation among pair members induced by matching. Lachin (1992b) describes a simpler direct calculation based on a specification of the marginal probability of exposure among the controls ($\phi_{1\bullet}$), the conditional odds ratio to be detected ($\varphi_c = \phi_{12}/\phi_{21}$), and the correlation of exposures within pairs represented by the exposure odds ratio in the matched sample, $\omega = (\phi_{11}\phi_{22})/(\phi_{12}\phi_{21})$. From these quantities the following joint relationships apply:

$$\phi_d = \phi_{12} + \phi_{21} \quad (5.79)$$

$$\phi_{21} = \phi_{12}/\varphi_c$$

$$\phi_{11} = \omega\phi_{12}\phi_{21}/\phi_{22}.$$

Noting that $\phi_{\bullet 1} = \phi_{11} + \phi_{21}$ and that $\phi_{22} = 1 - \phi_{\bullet 1} - \phi_{12}$, it can then be shown that $\phi_{\bullet 1}(1 - \phi_{\bullet 1})$ can be expressed as a function of ϕ_{12} alone such that

$$\frac{\omega - 1}{\varphi_c} \phi_{12}^2 + \frac{\varphi_c \phi_{\bullet 1} + (1 - \phi_{\bullet 1})}{\varphi_c} \phi_{12} - \phi_{\bullet 1}(1 - \phi_{\bullet 1}) = 0. \quad (5.80)$$

This is a standard quadratic equation of the form $a\phi_{12}^2 + b\phi_{12} + c = 0$, and the root in the interval (0,1) provides the value of ϕ_{12} that satisfies these constraints. Given ϕ_{12} , the other probabilities of the table are then obtained from the above relationships. These probabilities $\{\phi_{ij}\}$ in the four cells of the table, or equivalently $\{\pi_{ij}\}$, can then be used in the unconditional sample size calculation in (5.72).

Example 5.11 Case-Control Study

Lachin (1992b) presents the following example (reproduced with permission). We wish to determine the total sample size for a matched case-control study that will provide unconditional power to detect a conditional odds ratio $\varphi_c = 2$ with power $1 - \beta = 0.90$ in a two-sided test at level $\alpha = 0.05$. If the probability of exposure among the controls is $\phi_{\bullet 1} = 0.3$, then Schlesselman's calculation (not shown) yields $N = 186.4$. This assumes, however, that matching is ineffective or that the exposure odds ratio is $\omega = 1$. If we assume that matching is effective, such that $\omega = 2.5$, then the root of (5.80) yields $\phi_{12} = 0.25$, which yields the following table of probabilities:

		\bar{D}			
		E	\bar{E}		
D	E	$\phi_{11} = 0.175$	$\phi_{12} = 0.250$	$\phi_{1\bullet} = 0.425$	
	\bar{E}	$\phi_{21} = 0.125$	$\phi_{22} = 0.450$	$\phi_{2\bullet} = 0.575$	
		$\phi_{\bullet 1} = 0.30$	$\phi_{\bullet 2} = 0.70$	1	

The unconditional sample size calculation in (5.72) then yields $N = 248$, substantially larger than that suggested by Schlesselman's procedure.

Note that the discordant probabilities are the same as those specified in Example 5.9, so that the required sample size is the same. The only difference is how the probabilities arose. In Example 5.9 the problem was uniquely specified in terms of the probabilities of the discordant pairs. In Example 5.11, these probabilities are derived indirectly, starting with the specification of one of the marginal probabilities, the conditional odds ratio, and the exposure odds ratio, from which the discordant probabilities are then obtained.

5.6.4 Matching Efficiency

Various authors have considered the topic of the relative efficiency of matching versus not. Miettinen (1968; 1970) argues that "matching tends to reduce efficiency and would therefore have to be motivated by the pursuit of validity alone." He argues that the matched study has less power than an unmatched study, all other things being equal. However, his comparison and conclusions are flawed. Karon and Kupper

(1982), Kupper et al. (1981), Wacholder and Weinberg (1982), and Thomas and Greenland (1983) present further explorations of the efficiency of matching. In general, they conclude that efficient matching which provides positive within-pair correlation will provide greater power in a matched study than in an unmatched study of the same size.

5.7 STRATIFIED ANALYSIS OF PAIR-MATCHED TABLES

In some instances it may be desirable to conduct an analysis of pair-matched data using stratification or a regression model to adjust for other characteristics that may be associated either with exposure or the response. There have been few papers on stratification adjustment, in part because the conditional logistic regression model provides a convenient mechanism to conduct such an adjusted analysis. This model is described in Chapter 7. Nevertheless, it is instructive to consider a stratified analysis of K independent 2×2 tables that have been either prospectively or retrospectively matched in parallel with the methods developed for independent observations in Chapter 4. Here we only consider the case of matched pairs. In such a stratified matched analysis, we must distinguish between two different cases: pair versus member stratification.

5.7.1 Pair and Member Stratification

In *pair stratification*, the N matched pairs are characterized by pair-level covariates, where both members of the pair share the same values of the stratifying covariates. The pairs and their respective members then are grouped into K independent strata. Since all pair members have the same covariate values within each stratum, there is no covariate by exposure group association, and the marginal, unadjusted odds ratio is unbiased. Thus, some authors have stated that a stratified analysis in this case is unnecessary because the exposed and nonexposed groups are also balanced with respect to the pair stratification covariates. This, however, ignores the possible association between the stratifying covariates and the response, for which a stratification adjustment may be desirable (see Section 4.4.3).

The most common example of pair stratification is when the matched observations are obtained from the same individual, such as right eye versus left eye (one treated topically, the other not), or when observations are obtained before and again after some exposure, as in Example 5.7. In such cases, any patient characteristic is a potential pair-stratifying covariate. With such pair-matched data, it might still be relevant to conduct an analysis stratifying on potential risk factors that may be associated with the probability of a positive outcome (D), including the matching covariates. As we shall see, such pair stratification has no effect in a Mantel-Haenszel analysis, but it does yield an adjusted *MVLE* of the common odds ratio. If, for example, a study employed before-and-after or left-and-right measurements in the same individual, and if gender were a potent risk factor, then the gender stratified-adjusted *MVLE* of the odds ratio might be preferred. In such cases it would also be

important to assess the heterogeneity of the odds ratios across strata using a test of homogeneity.

In *member stratification* the individual members of each pair are stratified with respect to the covariate characteristics of the pair members. For example, consider a study using pairs of individuals that are matched with respect to age and severity of disease, but not by gender. It might then be desirable to conduct a stratified analysis that adjusts for gender. In this case, there are four possible strata with respect to the gender of the pair members: MM, MF, FM, and FF. Usually, the stratified analysis is then based on those concordant pairs where the two pair members are concordant with respect to the stratifying covariates: that is, using only the MM and FF strata. This provides an estimate of the conditional odds ratio in the setting where the members are also matched for gender in addition to the other matching covariates. Clearly, this sacrifices any information provided by the other strata wherein the pair members are discordant for the stratifying covariates: that is, the MF, FM strata. While these discordant strata would exist in the general unmatched population, their exclusion here is appropriate since the frame of reference is the matched population.

5.7.2 Stratified Mantel-Haenszel Analysis

Consider either pair stratification with K separate strata defined by the covariate values, or member stratification with K strata defined from the possible concordant covariate values for the two members of each pair. Building on the construction for a single matched 2×2 table shown in Table 5.2, there are a total of $4K$ sets of tables of type e_k , f_k , g_k , and h_k ($k = 1, \dots, K$), as presented in Sections 5.4.4 and 5.4.5. Applying the expression for the Mantel-Haenszel estimate of the common odds ratio in (4.14) yields

$$\widehat{OR}_{C(MH)} = \frac{\sum_k f_k}{\sum_k g_k} = \frac{f_{\bullet}}{g_{\bullet}}, \quad (5.81)$$

where f_{\bullet} and g_{\bullet} are the numbers of discordant pairs in the aggregate matched 2×2 table. Likewise, the expression for the Robins et al. (1986) variance in (4.28) is

$$\widehat{V} \left[\log \left(\widehat{OR}_{C(MH)} \right) \right] = \frac{1}{f_{\bullet}} + \frac{1}{g_{\bullet}} = \frac{M_{\bullet}}{f_{\bullet} g_{\bullet}}, \quad (5.82)$$

where $M_{\bullet} = f_{\bullet} + g_{\bullet}$. Similarly, the Mantel-Haenszel estimate of the common relative risk in (4.17) is

$$\widehat{RR}_{A(MH)} = \frac{\sum_k (e_k + f_k)}{\sum_k (e_k + g_k)} = \frac{e_{\bullet} + f_{\bullet}}{e_{\bullet} + g_{\bullet}}. \quad (5.83)$$

This relative risk estimate may be used for a matched prospective or cross-sectional study, but not for a case-control study.

Applying the expression for the Mantel-Haenszel test in (4.8) to these $4K$ tables yields the Mantel-Haenszel test statistic

$$X_{MH}^2 = \frac{\left[\sum_k (f_k - g_k) \right]^2}{\sum_k (f_k + g_k)} = \frac{(f_{\bullet} - g_{\bullet})^2}{f_{\bullet} + g_{\bullet}}, \quad (5.84)$$

which is the unadjusted McNemar test statistic. Thus, the stratified-adjusted Mantel-Haenszel estimates and test are the same as the unadjusted estimates and test.

5.7.3 MVLE

As was the case for 2×2 tables of independent (unmatched) observations, the MVLE of the common log odds ratio θ is obtained as the weighted combination of the stratum-specific values $\widehat{\theta}_k = \log(\widehat{OR}_{Ck})$ of the form

$$\widehat{\theta} = \sum_k \widehat{\omega}_k \widehat{\theta}_k \quad (5.85)$$

with estimated weights obtained from the estimated variance of the estimate (5.53),

$$\widehat{\omega}_k = \frac{(f_k g_k)/M_k}{\sum_\ell (f_\ell g_\ell)/M_\ell}. \quad (5.86)$$

The variance is consistently estimated as

$$\widehat{V}(\widehat{\theta}) = \frac{1}{\sum_\ell (f_\ell g_\ell)/M_\ell}, \quad (5.87)$$

which provides asymmetric confidence limits for the adjusted odds ratio.

Likewise, the MVLE of the common marginal log relative risk, when such is appropriate, is obtained using $\widehat{\theta}_k = \log(\widehat{RR}_{Ak})$ with weights obtained from the variance of the estimate (5.60)

$$\widehat{\omega}_k = \frac{(n_{1\bullet k} n_{\bullet 1 k})/M_k}{\sum_\ell (n_{1\bullet \ell} n_{\bullet 1 \ell})/M_\ell}, \quad (5.88)$$

and large sample estimated variance

$$\widehat{V}(\widehat{\theta}) = \frac{1}{\sum_\ell (n_{1\bullet \ell} n_{\bullet 1 \ell})/M_\ell}. \quad (5.89)$$

These methods can be applied to either pair- or member-stratified analyses.

5.7.4 Tests of Homogeneity and Association

5.7.4.1 Conditional Odds Ratio The contrast or Cochran's test of homogeneity of the conditional odds ratios is readily conducted using the MVLE $\widehat{\theta}$ in (5.85) in conjunction with the inverse variances $\widehat{\tau}_k = \widehat{V}(\widehat{\theta}_k)^{-1}$ for each stratum ($k = 1, \dots, K$). These quantities are simply substituted in the expression for the test statistic in (4.67). Alternatively, Breslow and Day (1980) suggest that a simple $K \times 2$ contingency table of the discordant frequencies n_{1k} and n_{2k} could be constructed and used as the basis for a simple contingency chi-square test as in (2.84) on $K-1$ df . It is readily shown (Problem 5.12.11) that the null hypothesis of independence

in the $K \times 2$ table is equivalent to the hypothesis of homogeneity of the conditional odds ratio among strata.

The asymptotically efficient test of association of a common conditional odds ratio can be derived as follows. Since the conditional odds ratio is a function of the discordant probabilities, this suggests that the analogous Radhakrishna-like test statistic is of the form

$$T = \sum_{k=1}^K w_k (f_k - g_k). \quad (5.90)$$

Now assume that $E(f_k) = N_k \pi_{12k}$, $E(g_k) = N_k \pi_{21k}$, and that $\theta_k = g(\pi_{12k}) - g(\pi_{21k}) = \theta \forall k$, where $g(\pi) = \log(\pi)$. Then by a first-order Taylor's expansion, asymptotically

$$E(T) \cong \theta \sum_{k=1}^K \frac{w_k N_k}{g'(\pi_{dk}/2)} = \theta \sum_{k=1}^K \frac{w_k N_k \pi_{dk}}{2}, \quad (5.91)$$

where $\pi_k = \pi_{dk}/2 \in (\pi_{12k}, \pi_{21k})$ and $\pi_{dk} = \pi_{12k} + \pi_{21k}$. From (5.35), noting that $p_{12} = f/N$ and that $p_{21} = g/N$, the large sample variance of the statistic is

$$V(T | H_0) \cong \sum_{k=1}^K w_k^2 V(f_k - g_k | H_0) = \sum_{k=1}^K w_k^2 N_k \pi_{dk}. \quad (5.92)$$

Evaluating the asymptotic efficiency of the test, it follows that the optimal weights are $w_k = 1$ for all strata and that the asymptotically efficient test is the same as that derived as the Mantel-Haenszel test in (5.84).

Thus, for a pair- or a member- stratified analysis, the Radhakrishna-like asymptotically efficient test of a common odds ratio is the same as the marginal unadjusted McNemar test. This is not surprising because the Mantel-Haenszel test is a member of the Radhakrishna family.

5.7.4.2 Marginal Relative Risk Similarly, in a matched prospective study, a test of homogeneity of log relative risks is readily constructed using the MVLE and the variance of the estimates within strata as in (5.58) and (5.60).

Also, an asymptotically efficient test of a common relative risk can readily be derived. The log RR is $\theta_k = g(\pi_{1\bullet k}) - g(\pi_{\bullet 1 k})$, where $g(\pi) = \log(\pi)$, as above. Using the test statistic

$$T = \sum_{k=1}^K w_k (p_{1\bullet k} - p_{\bullet 1 k}), \quad (5.93)$$

then, under the model with a common relative risk,

$$E(T) \cong \theta \frac{\sum_{k=1}^K w_k}{g'(\pi_k)}, \quad (5.94)$$

where $\pi_k \in (\pi_{1\bullet k}, \pi_{\bullet 1 k})$. The null variance is

$$V(T | H_0) \cong \sum_{k=1}^K w_k^2 V(p_{1\bullet k} - p_{\bullet 1 k} | H_0) = \sum_{k=1}^K w_k^2 \pi_{dk} / N_k \quad (5.95)$$

since $p_{1\bullet k} - p_{\bullet 1k} = p_{12k} - p_{21k}$. Then it is easily shown that the asymptotically efficient test uses the optimal weights

$$w_k = \frac{\pi_k N_k}{\pi_{dk}} \hat{=} \frac{(p_{1\bullet k} + p_{\bullet 1k})N_k}{2(p_{12k} + p_{21k})}. \quad (5.96)$$

These derivations are left to Problem 5.12.10.

5.7.4.3 Robust Tests Since the optimal test of a common odds ratio and the test of a common relative risk employ different statistics, the *MERT* for the family comprising these two tests is not described. However, the correlation can be obtained by expressing each of these asymptotically efficient tests as a linear combination of the vectors of $4K$ frequencies or proportions.

Alternatively, the Wei-Lachin test of stochastic ordering could be applied of the form

$$X_S^2 = \frac{[\sum_k (p_{12k} - p_{21k})]^2}{\sum_k (p_{12k} + p_{21k})/N_k}, \quad (5.97)$$

that is based on the unweighted sum (or mean) of the discordant proportion difference among strata.

Example 5.12 *Pregnancy and Retinopathy Progression*

In the analysis of the effects of pregnancy on the level of retinopathy (see Example 5.7), it was also important to assess the influence of the effect of intensification of control of blood glucose levels on the level of retinopathy. All women who planned to become pregnant, or who became pregnant, received intensified care to achieve near-normal levels of blood glucose to protect the fetus. However, other studies had shown that such intensification alone could have a transient effect on the level of retinopathy that could be confounded with the effect of pregnancy. Thus, it was important to adjust for this effect in the analysis. Here the stratified analysis within the intensive treatment group is described. The analysis within the conventional treatment group is used as Problem 5.13.

The degree of intensification of therapy was reflected by the change in the level of HbA_{1c} from before to during pregnancy: the greater the reduction, the greater the intensification. The following table presents the numbers within each of four strata classified according to the decrease in HbA_{1c} during pregnancy, where a negative value represents an increase.

<i>HbA_{1c}</i> <i>Decrease</i>	<i>N</i>	<i>Better</i> # (%)	<i>Worse</i> # (%)	<i>OR_c</i>	95% <i>C.I.</i>
<0	15	5 (33)	6 (40)	1.2	0.4, 3.9
(0.0 %, 0.7 %]	20	5 (25)	10 (50)	2.0	0.7, 5.9
(0.7 %, 1.3 %]	22	1 (5)	6 (27)	6.0	0.7, 50.0
>1.3 %	20	3 (15)	9 (45)	3.0	0.8, 11.1

(DCCT, 2000, reproduced with permission). The *MVLE* provides a stratified-adjusted odds ratio of 2.1 with 95% confidence limits of (1.10, 4.02), the estimate being nearly

identical to the unadjusted value of 2.2 presented in Example 5.7. A Wald test based on the estimated adjusted log odds ratio (0.743) and its estimated variance (0.109) yields $X^2 = (0.743)^2/0.109 = 5.067$, which is also highly significant ($p < 0.006$). This test value is comparable to that provided by the unadjusted McNemar test in Example 5.7.

The Cochran test of homogeneity of odds ratios over strata yields $X^2 = 2.093$ on 3 df with $p \leq 0.56$. A simpler test that does not require the computation of the *MVLE* is to conduct a Pearson contingency test of independence for the above 4×2 table of frequencies, for which $X^2 = 2.224$ on 3 df with $p \leq 0.527$. Both tests indicate that the effect of pregnancy on the level of retinopathy does not depend on the change in the level of HbA_{1c} during pregnancy.

The Wei-Lachin tests of stochastic ordering would be appropriate when it is desired to detect a difference in the same direction over strata. This results in $X_S^2 = 6.66$ on 1 df with $p < 0.01$. This test, if prespecified, would be appropriate even if there were heterogeneity among the strata.

Example 5.13 Member-Stratified Matched Analysis

To illustrate a member-stratified analysis, consider the endometrial cancer data of Example 5.8. One of the covariates measured on each participant in the study is the presence or absence of a history of hypertension (H vs. \bar{H}). Thus, there are four possible strata defined by the covariate values of the pair members: $\bar{H}\bar{H}$ with 16 discordant pairs, $\bar{H}H$ with seven pairs, $H\bar{H}$ with 11 pairs, and HH with five pairs. Within the concordant strata $\bar{H}\bar{H}$ and HH the numbers of discordant pairs are too small to provide a reliable analysis, $g_k = 3$ and 0 in each stratum, respectively. With a larger sample size the same analyses as in the Example 5.12 could then be performed using the concordant strata.

5.7.5 Random Effects Model Analysis

Based on the test of homogeneity, a random effects analysis of odds ratios, or relative risks when appropriate, is readily implemented as described in Section 4.10.2. The expressions previously presented apply, the only difference being the computational expressions for the log odds ratio (or relative risk) and the estimated variance within each stratum. Specifically, the test of homogeneity provides a moment estimator of the variance between strata, σ_θ^2 , based on (4.154). Then the total variance of each stratum-specific estimate involves the estimation variance plus the variance between strata, $V(\hat{\theta}_k) = V(\hat{\theta}_k|\theta_k) + \sigma_\theta^2$, as in (4.155). This then leads to a first-step reweighted estimate using the weights as in (4.156).

5.8 MULTIPLE MATCHING: MANTEL-HAENSZEL ANALYSIS

The prior sections all deal with pair or 1:1 matching, such as when a single control is matched to each case. Miettinen (1969) presents a test of association for $1:m_2$ matching with $m_2 > 1$ controls matched to each case. Walter (1980) presents a

further generalization for variable numbers of controls matched to each case; i.e., where $m_{2i} \in (1, 2, 3, \dots)$ for the i th matched set. Analyses are also conveniently obtained from a conditional logistic regression model as described in Chapters 6 and 7. In the most general case of N matched sets with possibly variable numbers of cases $\{m_{1i}\}$ and of controls $\{m_{2i} = n_i - m_{1i}\}$, or positive and negative responses, a Mantel-Haenszel analysis also provides both a test of association and an estimate of the conditional odds ratio that is readily computed using SAS or other software.

The i th matched set ($i = 1, \dots, N$) consists of n_i members, of whom $m_{1i} \geq 1$ are cases and $m_{2i} \geq 1$ are controls. As in (5.56), a 2×2 table is then constructed for each matched set of the form

	E	\bar{E}	
D	a_i	b_i	m_{1i}
\bar{D}	c_i	d_i	m_{2i}
n_{i+}	$n_i - n_{i+}$	n_i	

(5.98)

where n_{i+} is the number exposed in the i th set. The N matched sets then comprise N independent strata from which the aggregate Mantel-Haenszel test and estimated odds ratio can be computed, such as using SAS PROC FREQ.

This approach can also be used to conduct a stratified analysis over K strata wherein a 2×2 table is constructed for the matched sets separately within each stratum. The Mantel-Haenszel analysis is then based on the sum over all matched sets within all strata.

Example 5.14 Low Birth Weight

Hosmer and Lemeshow (2004) present data from a study of factors associated with the risk of giving birth to a low-weight infant. The complete data set is described in Chapter 7. Sets of three to 18 women who had a low (case) versus normal (control) birth weight delivery were matched by age. Among the risk factors evaluated is whether or not the mother smoked (the exposure variable). The following table presents the numbers of exposed and nonexposed cases and controls in the first eight matched sets (selected for convenience of display).

<i>Matched Set</i>	D	\bar{D}	E	\bar{E}	n_i	<i>Matched Set</i>	D	\bar{D}	E	\bar{E}	n_i
1	0	1	2	4	7	5	5	3	3	7	18
2	3	2	2	5	12	6	2	3	3	4	12
3	1	1	6	2	10	7	2	0	3	8	13
4	2	1	6	7	13	8	3	2	1	7	13

The Mantel-Haenszel procedure applied to the eight resulting 2×2 tables yields a test of association $X_{MH}^2 = 4.51$ with $p = 0.033$. The Mantel-Haenszel estimate of the conditional odds ratio is $\widehat{OR}_{MH} = 2.60$ with 95% confidence limits (1.057, 6.38). Because there is a zero in a cell in the 2×2 tables for strata 1 and 7, PROC FREQ

adds a value 0.5 to every cell of those two tables. As a result, the MVLE (logit) estimate is 2.52 with 95% limits (0.994, 6.38) that no longer bracket zero.

5.9 MATCHED POLYCHOTOMOUS DATA

The prior results also generalize to the analysis of matched polychotomous data. Let Y_1 and Y_2 denote the responses with C categories for the members of each of N matched pairs. Thus, the data can be represented as a $C \times C$ table with frequencies n_{ij} being the number of pairs in the i th row and j th column for the matched responses, and $p_{ij} = n_{ij}/N$ the corresponding proportion with expectation π_{ij} . Thus, the table frequencies are distributed as multinomial with C^2 parameters $\{\pi_{ij}\}$.

5.9.1 McNemar's Test

Example 5.7 presents one potential analysis of interest in which the discordant frequencies in symmetric upper and lower diagonal cells are pooled and the McNemar test conducted of the hypothesis that the pooled probability of falling in the upper area equals that in the lower area. For example, if the extreme three cells were used as the basis for the test in a 4×4 table, this approach would test the hypothesis that $\pi_{14} + \pi_{34} + \pi_{24} = \pi_{41} + \pi_{43} + \pi_{42}$. It would not test that the three opposite elements are equal; i.e., that $\pi_{14} = \pi_{41}$, $\pi_{24} = \pi_{42}$, and $\pi_{34} = \pi_{43}$.

5.9.2 Bowker's Test of Symmetry

Bowker (1948) presented a generalization of the McNemar test to address the hypothesis of elementwise symmetry: $H_0: \pi_{ij} = \pi_{ji}$ for all $i \neq j$ that the symmetric entries are equal. For a given pair, from the derivation of the McNemar test in Section 5.3.2, it follows that the equivalent of McNemar's test for a specific symmetric pair of cells is

$$X_{ij}^2 = \frac{(p_{ij} - p_{ji})^2}{(p_{ij} + p_{ji})/N} \quad \text{for } 1 \leq i < j \leq C, \quad (5.99)$$

that is distributed as chi-square on 1 df . It can then be readily shown that

$$\begin{aligned} \text{Cov}[(p_{ij} - p_{ji}), (p_{ik} - p_{ki})] &= 0, \quad i \neq j \neq k \\ \text{Cov}[(p_{ij} - p_{ji}), (p_{k\ell} - p_{\ell k})] &= 0, \quad i \neq j \neq k \neq \ell. \end{aligned} \quad (5.100)$$

It then follows that

$$X_B^2 = \sum_{i=1}^{C-1} \sum_{j=(i+1)}^C \frac{(p_{ij} - p_{ji})^2}{(p_{ij} + p_{ji})/N} \quad (5.101)$$

is the sum of independent 1 df chi-square statistics and thus is distributed as chi-square on $C(C - 1)/2$ df , (see Problem 5.14). This test is provided by SAS PROC FREQ with the *agree* option and is labeled the *test of symmetry*.

Bowker's test could also be used to test the hypothesis of elementwise symmetry within an upper and lower triangular area of the $C \times C$ table as in Example 5.7.

5.9.3 Marginal Homogeneity and Quasi-symmetry

Another hypothesis of interest is that of marginal homogeneity, which specifies that the row and column marginal probabilities are equal, or $H_0: \pi_{i\bullet} = \pi_{\bullet i}$ for $i = 1, \dots, C$. Cochran (1950) proposed his Q statistic to test this hypothesis for the set of margins in K 2×2 tables and this test is provided by the SAS PROC FREQ. Stuart (1955) and Bhapkar (1979), among others, have proposed tests of this hypothesis for $C \times C$ tables, and others have proposed generalizations to K such tables. Sun and Yang (2008) present a SAS macro that computes the Stuart and Bhapkar tests. Grizzle et al. (1969), among others, also show that a test of marginal homogeneity can be conducted using a log-linear model for contingency tables that can be applied using the SAS PROC CATMOD. White et al. (1982) provide a review of these and other methods.

When the marginal probabilities are heterogeneous, the hypothesis of quasi-symmetry specifies that there is symmetry between the upper and lower table elements after accounting for the difference in marginal probabilities (proportions). Agresti (1990) and Bishop et al. (1975) describe these methods in greater detail.

5.10 KAPPA INDEX OF AGREEMENT

5.10.1 Duplicate Gradings, Binary Characteristic

For pair-matched data, McNemar's test of symmetry, and the conditional odds ratio, assess the extent to which the proportions positive differ for the pair-matched variables. In some settings, such as duplicate gradings for the presence or absence of a characteristic, the objective is to describe and quantify the extent of agreement between the two gradings. The most commonly used index, among many, is the agreement statistic Kappa (κ), originally proposed by Cohen (1960) for binary paired assessments, and later generalized by Cohen (1968) to paired polychotomous assessments. Fleiss (1981) and Kraemer et al. (2004), among others, present comprehensive reviews of these methods.

Consider simple duplicate binary measures (Y_1, Y_2) obtained for each sample unit, such as two independent gradings of the presence of abnormality or disease for a given subject, often designated as the primary and secondary gradings. For a sample of N such paired gradings, the data can be expressed in a 2×2 table of frequencies with corresponding expectations.

		Frequencies		Probabilities	
		Secondary		Secondary	
Primary	$y_{i2} = 1$	$y_{i2} = 1$	$y_{i2} = 0$	$\pi_{1\bullet}$	$\pi_{2\bullet}$
		n_{11}	n_{12}		
$y_{i1} = 1$	$y_{i1} = 0$	n_{21}	n_{22}	π_{21}	π_{22}
		$n_{\bullet 1}$	$n_{\bullet 2}$	$\pi_{\bullet 1}$	$\pi_{\bullet 2}$
					1

The frequencies $(n_{11}, n_{12}, n_{21}, n_{22})$ are assumed to be distributed as multinomial with probabilities $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ that are estimated from the corresponding proportions $(p_{11}, p_{12}, p_{21}, p_{22})$, where $p_{ij} = n_{ij}/N$. The observed proportion of agreement is simply $p_O = p_{11} + p_{22}$. Given the marginal frequencies, the expected proportion agreement by chance is $p_E = p_{1\bullet}p_{\bullet 1} + p_{2\bullet}p_{\bullet 2}$. Then the estimated excess proportion agreement beyond chance, as a fraction of that possible beyond chance, is

$$\hat{\kappa} = \frac{p_O - p_E}{1 - p_E}, \quad \frac{-p_E}{1 - p_E} \leq \hat{\kappa} \leq 1. \quad (5.102)$$

Note that it is possible for the data to show less agreement than expected by chance, designated by a value $\hat{\kappa} < 0$.

The estimate of the large sample variance has been described by many (see Fleiss, 1981). To do so, it is convenient to express $\hat{\kappa}$ as a function of the cell frequencies. Expanding, it can be shown that

$$\begin{aligned} p_O - p_E &= p_{11} + p_{22} - p_E \\ &= 2p_{11}p_{22} - 2p_{12}p_{21} \\ 1 - p_E &= p_O - p_E + (p_{12} + p_{21}), \end{aligned} \quad (5.103)$$

so that the statistic can be expressed as

$$\hat{\kappa} = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} - p_{12}p_{21} + (p_{12} + p_{21})/2}. \quad (5.104)$$

The vector of proportions is distributed as multivariate normal with covariance matrix $\Sigma(\pi)/N$ as shown in Example A.2, where the subscripts $(11, 12, 21, 22)$ are equivalent to $(1, 2, 3, 4)$. Solving the vector of partial derivatives (\mathbf{H}) with elements $\{\partial\kappa/\partial\pi_{ij}\}$, the variance of $\hat{\kappa}$ is obtained as

$$\begin{aligned} \hat{V}(\hat{\kappa}) &= \mathbf{H}'\Sigma(\hat{\pi})\mathbf{H}/n = \frac{1}{N(1 - p_E)^2} \left\{ (1 - \hat{\kappa})^2 \sum_{i \neq j} p_{ij}(p_{\bullet i} + p_{\bullet j})^2 \right. \\ &\quad \left. + \sum_{i=1}^2 p_{ii} [1 - (p_{i\bullet} + p_{\bullet i})(1 - \hat{\kappa})]^2 - [\hat{\kappa} - p_E(1 - \hat{\kappa})]^2 \right\} \end{aligned} \quad (5.105)$$

(Fleiss et al., 1969). This is the unrestricted variance computed under the alternative that provides the estimated standard error and confidence limits. A simpler expression can be obtained if one assumes that the marginal probabilities of the gradings are equal (i.e., $\pi_{1\bullet} = \pi_{\bullet 1}$), as in Bloch and Kraemer (1989), who describe Kappa as an intraclass correlation (intraclass Kappa). Evaluating the variance under the null hypothesis of no excess agreement, or that $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$, yields the null variance estimate

$$\hat{V}_0(\hat{\kappa}) = \frac{p_E + p_E^2 - \sum_{i=1}^2 p_{i\bullet}p_{\bullet i}(p_{i\bullet} + p_{\bullet i})}{N(1 - p_E)^2} \quad (5.106)$$

(Fleiss et al. 2003) that forms the basis for a statistical test. These computations are provided by SAS PROC FREQ with the *agree* option.

Example 5.15 A Hypothetical Example

Consider the following two tables

A		Secondary		100
		1	0	
Primary	1	38	10	48
	0	30	22	52
	68	32		100

B		Secondary		100
		1	0	
Primary	1	88	5	93
	0	3	4	7
	91	9		100

In Table A the first grader reads 48% positive versus 68% for the second grader. The McNemar test of marginal homogeneity (symmetry) is thus highly significant ($p = 0.002$). Despite this, the proportion observed agreement ($p_O = 0.60$) exceeds the chance agreement ($p_E = 0.493$), yielding $\hat{\kappa} = 0.211$ with variance $\hat{V}(\hat{\kappa}) = (0.0893)^2$, that yields 95% confidence limits (0.036, 0.387). The variance under the null is $\hat{V}_0(\hat{\kappa}) = (0.092)^2$, that yields $Z = 2.30$ and two-sided $p = 0.022$. Thus, as in this example, it is possible to observe both significant agreement and significant discordance (marginal heterogeneity) in the types of disagreements.

In Table B there is a very high proportion of agreement ($p_O = 0.92$). However, both gradings have a high proportion positive, that in turn yields a high proportion of agreement expected by chance ($p_E = 0.853$). As a result, $\hat{\kappa} = 0.46$. While still statistically significant, it is far less than the observed proportion agreement. This illustrates the anomaly that as $\pi_{1\bullet}$ and $\pi_{\bullet 1}$ approach 1, $\hat{\kappa}$ approaches zero.

5.10.2 Duplicate Gradings, Polychotomous or Ordinal Characteristic

For duplicate polychotomous gradings, the data comprise a $C \times C$ table as described previously for polychotomous matched assessments. The Kappa statistic can then be computed as in (5.102) with variances as in (5.105) and (5.106), where the summations are over 1 to C rather than 2. This computation is invariant to the ordering of the categories and thus is appropriate for a nominally scaled variable.

Cohen (1968) also proposed a weighted Kappa to describe agreement for duplicate ordinal gradings with a specific ordering represented by the scores assigned to the C categories (s_1, \dots, s_C) as are used in the mean score test or the test for trend described in Section 2.10. For the $C \times C$ matrix of paired frequencies, a symmetric table of weights $\{w_{ij}\}$ are assigned to each cell, $0 \leq w_{ij} \leq 1$, where $w_{ij} = w_{ji}$ and the diagonal entries are $w_{ii} = 1$. Cicchetti and Allison (1971) proposed a set of weights $\{w_{ij(CA)}\}$ using absolute differences in the corresponding marginal scores, whereas Fleiss and Cohen (1973) weights $\{w_{ij(FC)}\}$ use squared differences:

$$w_{ij(CA)} = 1 - \frac{|s_i - s_j|}{s_C - s_1} \quad (5.107)$$

$$w_{ij(FC)} = 1 - \frac{(s_i - s_j)^2}{(s_C - s_1)^2}$$

Table 5.3 Duplicate masked readings of the severity of diabetic retinopathy in the Diabetes Control and Complications Trial.

First Reading	Second Reading											
	Normal	MA Only	Mild NPDR	Moderate NPDR	1	2	3	4	5	6	7	8
1. 10/10	24	6										
2. 20/<20	4	6	2									
3. 20/20		2	16									
4. 30/<30			2	20	2							
5. 30/30				2	2	2						
6. 40/<40			2	2	4			8	2			
7. 40/40									2	2	4	
8. 45/<45										2	2	

for $1 \leq i \leq j \leq C$. Then weighted Kappa ($\widehat{\kappa}_w$) is computed as in (5.102) using the quantities

$$p_{Ow} = \sum_{i=1}^C \sum_{j=1}^C w_{ij} p_{ij} \quad (5.108)$$

$$p_{Ew} = \sum_{i=1}^C \sum_{j=1}^C w_{ij} (p_{i\bullet w} p_{\bullet j w}).$$

Technically, $\widehat{\kappa}_w$ is not an index of agreement per se, but rather, an index of similarity between the first and second ordinal gradings. When $w_{ij} = 0$ for $i \neq j$, then $\widehat{\kappa}_w = \widehat{\kappa}$.

The large sample variance is

$$\widehat{V}(\widehat{\kappa}) = \frac{1}{N(1 - p_{Ew})^2} \left\{ \sum_{i=1}^C \sum_{j=1}^C p_{ij} [w_{ij} - (p_{i\bullet w} + p_{\bullet j w})(1 - \widehat{\kappa}_w)]^2 - [\widehat{\kappa}_w - p_{Ew}(1 - \widehat{\kappa}_w)]^2 \right\}, \quad (5.109)$$

where

$$p_{i\bullet w} = \sum_{j=1}^C p_{\bullet j} w_{ij} \quad (5.110)$$

$$p_{\bullet j w} = \sum_{i=1}^C p_{i\bullet} w_{ij}$$

(Fleiss et al., 1969).

Example 5.16 Retinopathy in the Diabetes Control and Complications Trial

Lachin (2004) presents duplicate masked ordinal gradings of the severity of retinopathy (intra-retinal abnormalities) in a sample of 120 patients from the Diabetes Control and Complications Trial (DCCT, 1993). Retinopathy severity in each eye was graded as normal (10), the presence of microaneurysms (MA) only (20), mild nonproliferative diabetic retinopathy (NPDR, 30), and different levels of moderate NPDR (40, 45).

The combined score for each subject is coded as the grade for the worse/better eye, where the score for the better eye is either equal to that of the worse eye (e.g., 30/30) or less than that of the worse eye (e.g., 30/<30). Table 5.3 presents the duplicate reading results.

There is perfect agreement in 67% of the duplicates that yields an unweighted $\hat{\kappa} = 0.6012$, with $S.E.(\hat{\kappa}) = 0.0494$ and 95% confidence interval (0.5043, 6981). However, this ignores the ordinal nature of the gradings. Using the default Cicchetti-Allison weights in SAS, the more general weighted estimate is $\hat{\kappa}_w = 0.8069$, with $S.E.(\hat{\kappa}) = 0.0307$ and 95% confidence interval (0.7467, 8670).

However, the primary outcome in the study was progression by three or more steps from the level at baseline, and thus it is relevant to describe the reproducibility with which a change of three or more steps could be differentiated. For a c -step change the corresponding weight matrix in the SAS computation would use elements

$$w_{k\ell} = \begin{cases} 1 & \text{if } |k - \ell| < c \\ 0 & \text{otherwise,} \end{cases} \quad (5.111)$$

where in this case $1 \leq k \leq \ell \leq 8$. For a ± 3 -step change, the observed agreement is 98.3% that yields $\hat{\kappa}_w = 0.919$, with $S.E.(\hat{\kappa}) = 0.0400$ and 95% confidence interval (0.841, 998). SAS PROC FREQ does not provide the option to specify such an arbitrary weight matrix. The program *DCCTKappa.sas* performs these computations using PROC IML.

5.10.3 Multiple Gradings, Intraclass Correlation

Fleiss (1971) and Davies and Fleiss (1982), among others, describe generalizations of Kappa to the case of multiple ($r > 2$) replicate polychotomous gradings or assessments, and Fleiss et al. (1979) do so for variable numbers of gradings per subject. Landis and Koch (1977), among others, describe the calculation of the intraclass correlation for the case of multiple gradings and show that it is approximately equivalent to Kappa. Fleiss et al. (2003) and Shoukri (2004) provide further details.

5.11 PROBLEMS

5.1 For an unmatched case-control study, do the following:

5.1.1. Show that the prospective odds ratio in (5.5) equals the retrospective odds ratio in (5.4).

5.1.2. Show that the null hypothesis of equal prospective probabilities $H_0: P(D|E) = P(D|\bar{E})$ is equivalent to the null hypothesis of equal retrospective probabilities $H_0: P(E|D) = P(E|\bar{D})$.

5.1.3. Also show that the alternative hypothesis of unequal prospective probabilities $H_1: P(D|E) \neq P(D|\bar{E})$ is equivalent to the hypothesis of unequal retrospective probabilities $H_1: P(E|D) \neq P(E|\bar{D})$.

5.1.4. For a given prevalence $P(D) = \delta$, derive the expression for RR presented in (5.14).

5.1.5. Under the rare disease assumption show that $RR_{\delta \downarrow 0} \doteq OR_{retro}$.

5.1.6. Also show that

$$RR_{(\phi_1 - \phi_2) \downarrow 0} \doteq OR_{retro}. \quad (5.112)$$

5.1.7. For $\alpha_1 = P(E)$ known, under the rare disease assumption show that

$$\text{logit}(\widehat{PAR}) = \log \frac{\widehat{PAR}}{1 - \widehat{PAR}} \doteq \log [\alpha_1(\widehat{OR} - 1)]. \quad (5.113)$$

5.1.8. Derive the expression for the variance of the $\text{logit}(\widehat{PAR})$ presented in (5.19).

5.2 Consider the following hypothetical unmatched case-control study of heart failure (D) and smoking (E).

	D	\overline{D}
E	120	130
\overline{E}	80	170
	200	300

5.2.1. Conduct an appropriate test of H_0 .

5.2.2. Calculate an estimate of the prospective population odds ratio with 95% confidence limits.

5.2.3. Assume that the prevalence of congestive heart failure is $P(D) = 1$ in 10,000. Calculate an estimate of the prospective population relative risk and compare to the estimate based on the odds ratio.

5.2.4. Calculate an estimate of the population attributable risk with 95% confidence limits assuming that $P(E) = \alpha_1 = 0.30$.

5.3 Consider matched pairs as described in Section 5.2.2.

5.3.1. For a binary response y_{ij} , where $E(y_{ij}|z_i) = \pi(z_i)$ for both members ($j = 1, 2$) of the i th pair, and where the responses (y_{i1}, y_{i2}) are conditionally independent, show that

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= \text{Cov}[E_z(y_{i1}|z), E_z(y_{i2}|z)] \\ &= E_z [E(y_{i1}|z_i)E(y_{i2}|z_i)] - E_z [E(y_{i1}|z_i)] E_z [E(y_{i2}|z_i)] \end{aligned} \quad (5.114)$$

and that

$$\text{Cov}(Y_1, Y_2) = E_Z [\pi(z)^2] - (E_Z [\pi(z)])^2 = V[\pi(z)] = \sigma_\pi^2 \neq 0. \quad (5.115)$$

5.3.2. Using the partitioning of variation as in (A.6), show that

$$V(Y_1) = V(Y_2) = E_Z \{\pi(z) [1 - \pi(z)]\} + \sigma_\pi^2. \quad (5.116)$$

5.3.3. Then show that

$$\text{Corr}(Y_1, Y_2) = \frac{\sigma_\pi^2}{E_Z \{\pi(z) [1 - \pi(z)]\} + \sigma_\pi^2}. \quad (5.117)$$

5.3.4. For a quantitative response assume that $y_{ij} = \mu(z_i) + \varepsilon_{ij}$, where $\mu(z_i) = E(y_{ij}|z_i)$, and where ε_{ij} are independent of μ_i with $E(\varepsilon_{ij}) = 0$, $V(\varepsilon_{ij}) = \sigma_\varepsilon^2 \forall ij$ and $\text{Cov}(\varepsilon_{i1}, \varepsilon_{i2}) = 0$. Then show that

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= E_z [\text{Cov}(y_{i1}, y_{i2}|z)] = E_z [\mu(z)^2] - (E_Z [\mu(z)])^2 \\ &= \sigma_\mu^2 \neq 0 \end{aligned} \quad (5.118)$$

and that

$$\text{Corr}(Y_1, Y_2) = \frac{E_Z [\mu(z)^2] - (E_Z [\mu(z)])^2}{\left(E_Z [\mu(z)^2] - (E_Z [\mu(z)])^2 \right) + \sigma_\varepsilon^2} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2}. \quad (5.119)$$

5.4 Consider a matched prospective study of recovery from depression (D) versus not (\bar{D}) among matched pairs of subjects, one of whom received a new treatment (E), the other a placebo (\bar{E})

			\bar{E}
		D	\bar{D}
E	D	16	3
\bar{D}	\bar{D}	7	4

5.4.1. State appropriate null and alternative hypotheses (H_0, H_1) and conduct an exact test of H_0 .

5.4.2. Calculate an estimate of the population odds ratio with exact 95% confidence limits.

5.5 Starting from the multinomial likelihood for a pair-matched study presented in (5.24), do the following:

5.5.1. Derive the conditional binomial distribution presented in (5.30).

5.5.2. Show that asymptotically

$$p_f = f/M \xrightarrow{d} \mathcal{N} \left(\frac{\pi_{12}}{\pi_d}, \frac{\pi_{12}\pi_{21}}{M\pi_d^2} \right). \quad (5.120)$$

5.6 For a pair-matched study, use the normal approximation to the multinomial as described in Section 5.3.2 to show the following:

5.6.1. Derive the expression for $V(p_{12} - p_{21})$ in (5.34) under the alternative hypothesis.

5.6.2. Then show that under the null hypothesis, $V(p_{12} - p_{21}|H_0) = \pi_d/N$, where $\pi_d = \pi_{12} + \pi_{21} = 2\pi$.

5.6.3. Also, show that the large sample Z -test is as shown in (5.36).

5.6.4. Show that $Z \xrightarrow{d} \mathcal{N}(0, 1)$ under H_0 .

5.6.5. For the conditional odds ratio defined in (5.43) and (5.46), let $\theta = \log(OR_C)$ and $\hat{\theta} = \log(f/g)$. Then use the δ -method to derive the expression for $V(\hat{\theta})$ in (5.52) and the consistent estimator in (5.53).

5.6.6. Likewise, for the log population averaged relative risk \widehat{RR}_A in (5.57), derive the expression for the variance in (5.59) and the estimate in (5.60).

5.7 Alternatively, one can base the test and the confidence limits on the conditional binomial distribution of f given M presented in (5.30) and derived in Problem 5.5.

5.7.1. Under the null hypothesis of symmetry $H_0: \pi_{12} = \pi_{21}$, show that $f \sim B(f; M, 1/2)$.

5.7.2. Show that this then implies that asymptotically

$$p_f \xrightarrow{d} \mathcal{N}\left(\frac{1}{2}, \frac{1}{4M}\right). \quad (5.121)$$

5.7.3. Then show that the resulting large sample Z -test for this hypothesis is

$$Z = \frac{f/M - 1/2}{\sqrt{1/(4M)}} \quad (5.122)$$

and that this equals McNemar's Z -test in (5.36).

5.7.4. Under $H_1: \pi_{12} \neq \pi_{21}$, from the large sample distribution of p_f derived in Problem 5.5.2 and shown in Section 5.4.3, show that the large sample confidence limits based on the logit of $\pi_f = \pi_{12}/\pi_d$ equal the confidence limits for $\theta = \log(OR_C)$ based on the estimated variance derived in Problem 5.6.5 and presented in (5.53).

5.8 Now consider a larger replication of the study in Problem 5.4 with hypothetical frequencies

		\overline{E}
	D	\overline{D}
E	D	135 18
	\overline{D}	62 24

5.8.1. State appropriate null and alternative hypotheses (H_0, H_1) and conduct a large sample test of H_0 .

5.8.2. Calculate an estimate of the conditional odds ratio with asymmetric 95% confidence limits.

5.8.3. Calculate an estimate of the population-averaged relative risk with asymmetric 95% confidence limits.

5.9 For the pair-matched 2×2 table, consider a Mantel-Haenszel analysis stratified by the N pairs as shown in Table 5.2 and described in Section 5.4.4.

5.9.1. Show that the a Mantel-Haenszel test statistic equals McNemar's test or that $X_{MH}^2 = X_M^2$ in (5.37).

5.9.2. Show that the Mantel-Haenszel estimate of the odds ratio equals the conditional odds ratio estimate, $\widehat{OR}_{MH} = f/g = \widehat{OR}_C$.

5.9.3. Show that the Robins et al. (1986) estimate of the variance of the log odds ratio in (4.28) equals that in (5.52) or that $V[\log \widehat{OR}_{MH}] = V[\log \widehat{OR}_C] \hat{=} M/(fg)$.

5.9.4. Show that the Mantel-Haenszel estimate of the relative risk equals the population-averaged RR_A in (5.58) or that $\widehat{RR}_{MH} = \widehat{RR}_A$.

5.10 Now consider the pair-matched retrospective study as described in Section 5.5.

5.10.1. Show that the hypothesis of symmetry with respect to the retrospective probabilities of discordant pairs implies and is implied by the hypothesis of symmetry for the prospective probabilities, that is, that $(\phi_{12|z} = \phi_{21|z}) \iff (\pi_{12|z} = \pi_{21|z})$.

5.10.2. Show that the retrospective conditional odds ratio equals the prospective conditional odds ratio as described in (5.66) and (5.67).

5.10.3. Derive the relation shown in (5.69) between the prospective conditional relative risk from a pair-matched case-control study and the retrospective conditional odds ratio.

5.10.4. In a retrospective study, from (5.69) show that $RR_z \doteq OR_C$ under the rare disease assumption ($\delta \downarrow 0$).

5.11 In a pair-matched study, from the distribution of $T = p_{12} - p_{21}$ under H_0 and H_1 in Section 5.6, do the following:

5.11.1. Derive the expression relating the total sample size and power in (5.72) unconditionally, that is, without fixing the number of discordant pairs, M .

5.11.2. From the unconditional power function of McNemar's test presented in (5.72) show that the equation for the N needed to provide power $1 - \beta$ for an α -level test is given by

$$N = \left[\frac{Z_{1-\alpha}\sqrt{\pi_d} + Z_{1-\beta}\sqrt{\pi_d - (\pi_{12} - \pi_{21})^2}}{\pi_{12} - \pi_{21}} \right]^2. \quad (5.123)$$

5.11.3. Then use the conditional distribution with p_f as the test statistic as described in Problem 5.5.2 to derive the expression relating sample size and power as

$$\sqrt{M} |\varphi_c - 1| = Z_{1-\alpha} (\varphi_c + 1) + 2Z_{1-\beta}\sqrt{\varphi_c}, \quad (5.124)$$

which is expressed as a function of M and φ_c alone.

5.11.4. From this, derive the expression for $Z_{1-\beta}$ in (5.77).

5.11.5. Derive the quadratic function in (5.80) given the parameters ϕ_{*1} , ω , and φ_c .

5.11.6. For a study where $\pi_{21} = 0.25$, $\alpha = 0.05$ (two-sided), what N is required to provide 90% power to detect an odds ratio of 1.8?

5.11.7. Show that the expected number of discordant pairs with this sample size is $E(M) = 126$.

5.11.8. Suppose that because of cost considerations, the final sample size is determined to be $N = 150$ with $E(M) = 105$.

1. From (5.77), what level of power does this provide to detect an odds ratio of 1.8?
2. What odds ratio can be detected with power = 0.90 with this sample size?

5.11.9. At the end of the study, however, assume that only $M = 90$ discordant pairs are observed. What level of power does this provide to detect an odds ratio of 1.8?

5.11.10. For a study where $\phi_{*1} = 0.2$, $\omega = 1.5$, and $\varphi_c = 1.3$, determine the cell probabilities $\{\phi_{ij}\}$ using the solution to (5.80). Then determine the sample size N required to provide 90% power to detect this value of φ_c .

5.12 Now consider a pair- or member-stratified analysis with K strata.

5.12.1. Show that a stratified Mantel-Haenszel analysis with $4K$ strata yields the adjusted estimate of the odds ratio and the estimated variance of the log odds ratio that equal those from the unadjusted marginal analysis as shown in (5.81)–(5.82).

5.12.2. Likewise, show that the stratified Mantel-Haenszel estimate of the relative risk equals that from the marginal unadjusted analysis shown in (5.83).

5.12.3. Also, show that the stratified Mantel-Haenszel test of a common odds ratio equals the marginal McNemar test as in (5.84).

5.12.4. Show that the *MVLE* of a common odds ratio uses the weights presented in (5.86) and has the variance presented in (5.87).

5.12.5. Likewise, show that the *MVLE* of a common relative risk uses the weights presented in (5.88) and has the estimated variance presented in (5.89).

5.12.6. For an asymptotically efficient test of a common odds ratio over strata, derive $E(T)$ in (5.91) with variance as shown in (5.92).

5.12.7. Then use the expressions for the asymptotic efficiency of the test as employed in Section 4.7.2 to show that the optimal weights that provide maximum efficiency for this test are $w_j = 1 \forall j$.

5.12.8. Then show that the resulting efficient *Z*-test equals the McNemar test in (5.36).

5.12.9. Similarly, for an asymptotically efficient test of a common relative risk over strata, derive $E(T)$ in (5.94) with variance as shown in (5.95).

5.12.10. Then show that the optimal weights $\{w_j\}$ that provide maximum efficiency for this test are as shown in (5.96).

5.12.11. Show that the hypothesis of homogeneity of the conditional odds ratio among strata is equivalent to the null hypothesis of independence in the $K \times 2$ contingency table.

5.13 Exercise 5.12 presents a stratified analysis of the effect of pregnancy on the level of retinopathy among women treated intensively in the Diabetes Control

and Complications Trial (DCCT, 2000). For women in the conventional treatment group, the following table presents the numbers of women whose retinopathy during pregnancy was better or worse than that before pregnancy within each of three strata classified according to the decrease in HbA_{1c} during pregnancy:

HbA_{1c} <i>Decrease</i>	<i>N</i>	<i>Better</i> # (%)	<i>Worse</i> # (%)	OR_c	95% <i>C.I.</i>
(0.0 %, 1.7%]	23	7 (30)	8 (35)	1.1	0.4, 3.2
(1.7 %, 3.1 %]	24	2 (8)	12 (50)	6.0	1.3, 26.8
>3.1 %	23	2 (9)	19 (82)	9.5	2.2, 40.8

The reductions in HbA_{1c} in this group were greater than those in the intensive group because the patients treated conventionally had higher levels of HbA_{1c} by about 2% during the study compared to the intensive group. Conduct the following analyses of these data.

5.13.1. In the unadjusted marginal analysis, a total of 11 patients had better and 39 worse levels of retinopathy during pregnancy than before pregnancy. Compute the conditional odds ratio and its asymmetric 95% confidence limits and conduct McNemar's test.

5.13.2. Compute the MVLE of the common log odds ratio across strata and its large sample variance. Use these to compute the 95% asymmetric confidence limits.

5.13.3. Then conduct Cochran's test of homogeneity of odds ratios over strata.

5.13.4. Also conduct a contingency chi-square test of independence for the 3×2 table of discordant frequencies.

5.13.5. Present an overall interpretation of the effect of pregnancy in the conventional treatment group on the level of retinopathy and the effect of the change in HbA_{1c} during pregnancy on this effect.

5.14 Consider Bowker's test for a $C \times C$ table as in Section 5.9.2.

5.14.1. Show that the covariances among elementwise differences in proportions equal 0 as in (5.100).

5.14.2. Then show that the vector of $C(C - 1)/2$ such differences is distributed as multivariate normal with a covariance matrix with elements $\Sigma = \text{diag}\{(p_{ij} + p_{ji})/N\}$.

5.14.3. Then show that Bowker's test in (5.101) is the sum of $C(C - 1)/2$ independent χ_1^2 variables.

5.15 Show that the expression for $\hat{\kappa}$ in (5.102) can be simplified as in (5.104).

Applications of Maximum Likelihood and Efficient Scores

Previous chapters describe the derivation of estimators and tests for the assessment of the association between exposure or treatment groups and the risk of an outcome from either independent or matched 2×2 tables obtained from either cross-sectional, prospective, and retrospective studies. Methods for the stratified analysis of multiple independent tables under fixed and random effects models are also described. All of those developments are based on classical statistical theory, such as least squares, asymptotic efficiency, and the central limit theorem.

Although Fisher's landmark work in maximum likelihood (Fisher, 1922b, 1925) predated many of the original publications, such as those of Cochran (1954a) and Mantel and Haenszel (1959), it was not until the advent of modern computers that maximum likelihood solutions to multiparameter estimation problems became feasible. Subsequently, it was shown that many of the classic methods could be derived using likelihood-based efficient scores. This chapter describes the application of this "modern" likelihood theory to develop the methods presented in previous chapters. Subsequent chapters then use these likelihood-based methods to develop regression models and related procedures. Section A.6 presents a review of maximum likelihood estimation and efficient scores, and Section A.7 describes associated test statistics.

6.1 BINOMIAL

Again consider the simple binomial distribution that describes the risk of an outcome in the population. Assuming that the number of positive responses X in N observations (trials) is distributed as binomial with $P(x) = B(x; N, \pi)$, the likelihood

is

$$L(\pi) = \binom{N}{x} \pi^x (1 - \pi)^{N-x}, \quad (6.1)$$

and the log likelihood is

$$\ell(\pi) = \log \binom{N}{x} + x \log \pi + (N - x) \log (1 - \pi). \quad (6.2)$$

Thus, the total score is

$$U(\pi) = \frac{d\ell}{d\pi} = \frac{x}{\pi} - \frac{N - x}{1 - \pi} = \frac{x - N\pi}{\pi(1 - \pi)}. \quad (6.3)$$

When set equal to zero, the solution yields the *MLE*:

$$\hat{\pi} = x/N = p. \quad (6.4)$$

The observed information is

$$i(\pi) = \frac{-d^2\ell}{d\pi^2} = \frac{x}{\pi^2} + \frac{N - x}{(1 - \pi)^2}. \quad (6.5)$$

Since $E(x) = N\pi$, the expected information is

$$I(\pi) = E[i(\pi)] = \frac{E(x)}{\pi^2} + \frac{E(N - x)}{(1 - \pi)^2} = \frac{N}{\pi(1 - \pi)}. \quad (6.6)$$

Thus, the large sample variance of the *MLE* is

$$V(\hat{\pi}) = I(\pi)^{-1} = \frac{\pi(1 - \pi)}{N}. \quad (6.7)$$

The estimated observed information is $i(\pi)_{|\pi=\hat{\pi}} = i(\hat{\pi})$ and the estimated expected information is $I(\pi)_{|\pi=\hat{\pi}} = I(\hat{\pi})$. For the binomial the two are equivalent:

$$i(\hat{\pi}) = I(\hat{\pi}) = \frac{N}{\hat{\pi}(1 - \hat{\pi})}, \quad (6.8)$$

and the estimated large sample variance of the *MLE* is

$$\hat{V}(\hat{\pi}) = I(\hat{\pi})^{-1} = \frac{\hat{\pi}(1 - \hat{\pi})}{N}. \quad (6.9)$$

Efron and Hinkley (1978) show that it is often better to use the observed information to compute confidence intervals, whereas for tests of hypotheses it is customary to use the expected information. In many problems, such as this, the two estimates are identical.

Since the number of positive responses is the sum of N *i.i.d.* Bernoulli variables, $x = \sum_{i=1}^N y_i$ as described in Section 2.1.2, then asymptotically $U(\pi)$ is normally

distributed with mean zero and variance $I(\pi)$. Likewise, as shown in Section A.6.5, then asymptotically

$$\hat{\pi} \xrightarrow{d} \mathcal{N}[\pi, V(\hat{\pi})],$$

or more precisely, as n increases, then

$$\sqrt{n}\hat{\pi} \xrightarrow{d} \mathcal{N}[\sqrt{n}\pi, \pi(1-\pi)].$$

All of these results were previously obtained from basic principles in Section 2.1.2.

Then, from the invariance principle of Section A.6.8, it follows that $g(\hat{\pi})$ is the *MLE* of any function $g(\pi)$, under nominal regularity conditions on $g(\cdot)$. The asymptotic distribution of $g(\hat{\pi})$ is then provided by the δ -method and Slutsky's theorem as illustrated in Section 2.1.3 using the logit and $\log(-\log)$ transformations.

6.2 2×2 TABLE: PRODUCT BINOMIAL (UNCONDITIONALLY)

Now consider the analysis of the 2×2 table from a study involving two independent groups of subjects, either from a cross-sectional, prospective, or retrospective study, unmatched. As in Chapter 2, the analysis of such a table is described from the perspective of a prospective study. Equivalent results apply to an unmatched retrospective study.

6.2.1 MLEs And Their Asymptotic Distribution

Given two independent samples of n_1 and n_2 observations, of which we observe x_1 and x_2 positive responses, respectively, then as shown in (2.53), the likelihood is a product binomial $L(\pi_1, \pi_2) = B(x_1; n_1, \pi_1)B(x_2; n_2, \pi_2) = L_1(\pi_1)L_2(\pi_2)$ and the log likelihood is $\ell = \log L_1 + \log L_2 = \ell_1 + \ell_2$. Thus, $\boldsymbol{\theta} = (\pi_1 \ \pi_2)^T$ and the score vector is

$$\begin{aligned} \mathbf{U}(\boldsymbol{\theta}) &= [U(\boldsymbol{\theta})_{\pi_1} \quad U(\boldsymbol{\theta})_{\pi_2}]^T = \frac{\partial \ell}{\partial \boldsymbol{\theta}} = \left[\frac{\partial \ell}{\partial \pi_1} \quad \frac{\partial \ell}{\partial \pi_2} \right]^T \\ &= \left[\frac{d\ell_1}{d\pi_1} \quad \frac{d\ell_2}{d\pi_2} \right]^T = \left[\frac{x_1 - n_1\pi_1}{\pi_1(1-\pi_1)} \quad \frac{x_2 - n_2\pi_2}{\pi_2(1-\pi_2)} \right]^T. \end{aligned} \quad (6.10)$$

Therefore, the *MLE* $\hat{\boldsymbol{\theta}}$ is the vector of the *MLEs* for the two independent groups $\hat{\boldsymbol{\theta}} = [\hat{\pi}_1 \ \hat{\pi}_2]^T$. The Hessian matrix has diagonal elements

$$\frac{\partial^2 \ell}{\partial \pi_1^2} = \frac{d^2 \ell_1}{d\pi_1^2} = - \left[\frac{x_1}{\pi_1^2} + \frac{n_1 - x_1}{(1-\pi_1)^2} \right] = \mathbf{H}(\boldsymbol{\theta})_{\pi_1} \quad (6.11)$$

$$\frac{\partial^2 \ell}{\partial \pi_2^2} = \frac{d^2 \ell_2}{d\pi_2^2} = - \left[\frac{x_2}{\pi_2^2} + \frac{n_2 - x_2}{(1-\pi_2)^2} \right] = \mathbf{H}(\boldsymbol{\theta})_{\pi_2}$$

and off-diagonal elements $\partial^2 \ell / (\partial \pi_1 \partial \pi_2) = 0$, since no terms in ℓ involve both π_1 and π_2 . Thus,

$$\mathbf{I}(\boldsymbol{\theta}) = E[\mathbf{i}(\boldsymbol{\theta})] = E[-\mathbf{H}(\boldsymbol{\theta})] = \begin{bmatrix} \frac{n_1}{\pi_1(1-\pi_1)} & 0 \\ 0 & \frac{n_2}{\pi_2(1-\pi_2)} \end{bmatrix}. \quad (6.12)$$

Then the large sample variance of $\widehat{\boldsymbol{\theta}} = [\widehat{\pi}_1 \ \widehat{\pi}_2]^T$ is

$$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \mathbf{I}(\boldsymbol{\theta})^{-1} = \text{diag} \left[\frac{\pi_1(1-\pi_1)}{n_1} \quad \frac{\pi_2(1-\pi_2)}{n_2} \right], \quad (6.13)$$

which is consistently estimated as

$$\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}}) = \mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1} = \text{diag} \left[\frac{p_1(1-p_1)}{n_1} \quad \frac{p_2(1-p_2)}{n_2} \right]. \quad (6.14)$$

Therefore, asymptotically

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left[\begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}, \begin{pmatrix} \frac{\pi_1(1-\pi_1)}{n_1} & 0 \\ 0 & \frac{\pi_2(1-\pi_2)}{n_2} \end{pmatrix} \right]. \quad (6.15)$$

From the invariance principle (Section A.6.8), for any one-to-one function of π such as $\beta = g(\pi_1) - g(\pi_2)$, it again follows that the *MLE* is $\widehat{\beta} = g(\widehat{\pi}_1) - g(\widehat{\pi}_2)$ with large sample variance of the estimate obtained by the δ -method. Thus, the *MLE* of the $\log(OR)$ is the sample log odds ratio, $\log(\widehat{OR})$, and the *MLE* of the $\log(RR)$ is $\log(\widehat{RR})$. These estimates and the large sample variance of each are presented in Section 2.3.

6.2.2 Logit Model

Since the odds ratio plays a pivotal role in the analysis of categorical data, it is also instructive to describe a simple logit or logistic model for the association in a 2×2 table, as originally suggested by Cox (1958a). This model was introduced in Problem 2.13. In this model the logits of the probabilities within each group can be expressed as

$$\log \left(\frac{\pi_1}{1 - \pi_1} \right) = \alpha + \beta, \quad \log \left(\frac{\pi_2}{1 - \pi_2} \right) = \alpha. \quad (6.16)$$

The inverse functions that relate the probabilities to the parameters $\boldsymbol{\theta} = (\alpha \ \beta)^T$ are logistic functions of the form

$$\begin{aligned} \pi_1 &= \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}}, & 1 - \pi_1 &= \frac{1}{1 + e^{\alpha+\beta}} \\ \pi_2 &= \frac{e^\alpha}{1 + e^\alpha}, & 1 - \pi_2 &= \frac{1}{1 + e^\alpha} \end{aligned} \quad (6.17)$$

and the odds ratio is simply

$$\frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = e^\beta. \quad (6.18)$$

This is called a *logit link model* or a *logistic model*. Later, other link functions will be introduced.

When expressed in terms of the frequencies (a, b, c, d) in the 2×2 table as in (2.24), the product binomial likelihood is

$$L(\pi_1, \pi_2) \propto \pi_1^a (1 - \pi_1)^c \pi_2^b (1 - \pi_2)^d, \quad (6.19)$$

and the log likelihood is

$$\ell(\pi_1, \pi_2) = a \log(\pi_1) + c \log(1 - \pi_1) + b \log(\pi_2) + d \log(1 - \pi_2). \quad (6.20)$$

Expressing the probabilities in terms of the parameters α and β and in terms of the marginal totals (n_1, n_2, m_1, m_2) in the 2×2 table, then

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= a \log \left[\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} \right] + c \log \left[\frac{1}{1 + e^{\alpha+\beta}} \right] \\ &\quad + b \log \left[\frac{e^\alpha}{1 + e^\alpha} \right] + d \log \left[\frac{1}{1 + e^\alpha} \right] \\ &= m_1 \alpha + a \beta - n_1 \log(1 + e^{\alpha+\beta}) - n_2 \log(1 + e^\alpha). \end{aligned} \quad (6.21)$$

The resulting score vector has elements $\mathbf{U}(\boldsymbol{\theta}) = [U(\boldsymbol{\theta})_\alpha \ U(\boldsymbol{\theta})_\beta]^T$, where

$$\begin{aligned} U(\boldsymbol{\theta})_\alpha &= \frac{\partial \ell}{\partial \alpha} = m_1 - n_1 \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} - n_2 \frac{e^\alpha}{1 + e^\alpha} \\ &= m_1 - n_1 \pi_1 - n_2 \pi_2 \end{aligned} \quad (6.22)$$

and

$$U(\boldsymbol{\theta})_\beta = \frac{\partial \ell}{\partial \beta} = a - n_1 \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} = a - n_1 \pi_1. \quad (6.23)$$

Note that these quantities can be expressed interchangeably as functions of the conditional expectations (π_1, π_2) or as functions of the model parameters (α, β) .

The *MLEs* $(\hat{\alpha}, \hat{\beta})$ are then obtained by setting each respective score equation to zero and solving for the parameter. This yields two equations in two unknowns. Taking the difference $U(\boldsymbol{\theta})_\alpha - U(\boldsymbol{\theta})_\beta$ then yields

$$\pi_2 = \frac{m_1 - a}{n_2}. \quad (6.24)$$

From (6.17), this implies that

$$\frac{e^\alpha}{(1 + e^\alpha)} = \frac{b}{n_2}. \quad (6.25)$$

Solving for α then yields

$$e^{\hat{\alpha}} = b / (n_2 - b) = b/d, \quad (6.26)$$

so that the *MLE* is $\hat{\alpha} = \log(b/d)$.

Then, evaluating $U(\boldsymbol{\theta})_\beta$ in (6.23) at the value $\hat{\alpha}$ and setting $U(\boldsymbol{\theta})_{\beta|\hat{\alpha}} = 0$ implies that

$$a - n_1 \frac{e^{\hat{\alpha}+\beta}}{1 + e^{\hat{\alpha}+\beta}} = 0. \quad (6.27)$$

Substituting the solution for $e^{\hat{\alpha}}$ yields

$$e^{\hat{\beta}} = \frac{a}{(n_1 - a) b/d} = \frac{ad}{bc}. \quad (6.28)$$

Therefore, the *MLE* of β is the sample log odds ratio $\hat{\beta} = \log(ad/bc)$, which was demonstrated earlier from the invariance principle.

The elements of the Hessian matrix then are

$$\mathbf{H}(\boldsymbol{\theta})_\alpha = \frac{\partial^2 \ell}{\partial \alpha^2} \quad (6.29)$$

$$= -n_1 \left[\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} - \left(\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} \right)^2 \right] - n_2 \left[\frac{e^\alpha}{1 + e^\alpha} - \left(\frac{e^\alpha}{1 + e^\alpha} \right)^2 \right]$$

$$= -n_1 [\pi_1 (1 - \pi_1)] - n_2 [\pi_2 (1 - \pi_2)]$$

$$\mathbf{H}(\boldsymbol{\theta})_\beta = \frac{\partial^2 \ell}{\partial \beta^2} = -n_1 \left[\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} - \left(\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} \right)^2 \right] = -n_1 [\pi_1 (1 - \pi_1)]$$

$$\mathbf{H}(\boldsymbol{\theta})_{\alpha\beta} = \frac{\partial^2 \ell}{\partial \alpha \partial \beta} = -n_1 [\pi_1 (1 - \pi_1)].$$

Thus, the observed information matrix is $\mathbf{i}(\boldsymbol{\theta}) = -\mathbf{H}(\boldsymbol{\theta})$, and the expected information matrix is

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta}) = \mathbf{i}(\boldsymbol{\theta}) &= \begin{bmatrix} n_1 [\pi_1 (1 - \pi_1)] + n_2 [\pi_2 (1 - \pi_2)] & n_1 [\pi_1 (1 - \pi_1)] \\ n_1 [\pi_1 (1 - \pi_1)] & n_1 [\pi_1 (1 - \pi_1)] \end{bmatrix} \\ &= \begin{bmatrix} n_1 \psi_1 + n_2 \psi_2 & n_1 \psi_1 \\ n_1 \psi_1 & n_1 \psi_1 \end{bmatrix}, \end{aligned} \quad (6.30)$$

where $\psi_i = \pi_i (1 - \pi_i)$ is the variance of the Bernoulli variate in the i th population, $i = 1, 2$. The elements of $\mathbf{I}(\boldsymbol{\theta})$ will be expressed as

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{I}(\boldsymbol{\theta})_\alpha & \mathbf{I}(\boldsymbol{\theta})_{\alpha,\beta} \\ \mathbf{I}(\boldsymbol{\theta})_{\alpha,\beta}^T & \mathbf{I}(\boldsymbol{\theta})_\beta \end{bmatrix}. \quad (6.31)$$

The covariance matrix of the estimates then is $\mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\boldsymbol{\theta})^{-1}$. The inverse of a 2×2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad (6.32)$$

is obtained as

$$A^{-1} = \begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix} \Big/ |A|, \quad (6.33)$$

where the determinant is $|A| = a_{11}a_{22} - a_{12}a_{21}$.

Therefore, $|\mathbf{I}(\boldsymbol{\theta})| = (n_1\psi_1 + n_2\psi_2)(n_1\psi_1) - (n_1\psi_1)^2 = n_1\psi_1 n_2\psi_2$ and the inverse matrix is

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta})^{-1} &= \frac{1}{n_1\psi_1 n_2\psi_2} \begin{bmatrix} n_1\psi_1 & -n_1\psi_1 \\ -n_1\psi_1 & (n_1\psi_1 + n_2\psi_2) \end{bmatrix} \quad (6.34) \\ &= \begin{bmatrix} \frac{1}{n_2\psi_2} & -\frac{1}{n_2\psi_2} \\ -\frac{1}{n_2\psi_2} & \frac{n_1\psi_1 + n_2\psi_2}{n_1\psi_1 n_2\psi_2} \end{bmatrix}. \end{aligned}$$

The elements of $\mathbf{I}(\boldsymbol{\theta})^{-1}$ will alternatively be expressed as

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta})^{-1} &= \begin{bmatrix} \left[\mathbf{I}(\boldsymbol{\theta})^{-1} \right]_{\alpha} & \left[\mathbf{I}(\boldsymbol{\theta})^{-1} \right]_{\alpha, \beta} \\ \left[\mathbf{I}(\boldsymbol{\theta})^{-1} \right]_{\alpha, \beta}^T & \left[\mathbf{I}(\boldsymbol{\theta})^{-1} \right]_{\beta} \end{bmatrix} \quad (6.35) \\ &= \begin{bmatrix} \mathbf{I}(\boldsymbol{\theta})^{\alpha} & \mathbf{I}(\boldsymbol{\theta})^{\alpha, \beta} \\ \mathbf{I}(\boldsymbol{\theta})^{\beta, \alpha} & \mathbf{I}(\boldsymbol{\theta})^{\beta} \end{bmatrix}. \end{aligned}$$

Thus, the elements of the covariance matrix $\mathbf{V}(\hat{\boldsymbol{\theta}})$ are

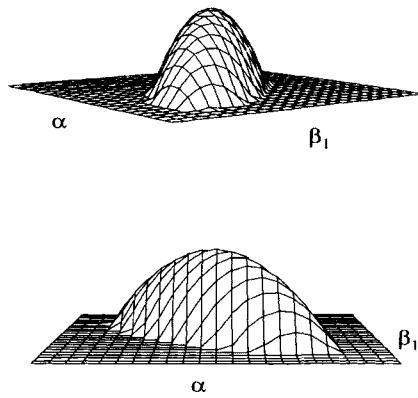
$$\begin{aligned} V(\hat{\alpha}) &= \left[\mathbf{I}(\boldsymbol{\theta})^{-1} \right]_{\alpha} = \frac{1}{n_2\pi_2(1-\pi_2)} \quad (6.36) \\ V(\hat{\beta}) &= \left[\mathbf{I}(\boldsymbol{\theta})^{-1} \right]_{\beta} = \frac{n_1\pi_1(1-\pi_1) + n_2\pi_2(1-\pi_2)}{n_1\pi_1(1-\pi_1)n_2\pi_2(1-\pi_2)} \\ &= \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)} \\ Cov(\hat{\alpha}, \hat{\beta}) &= \left[\mathbf{I}(\boldsymbol{\theta})^{-1} \right]_{\alpha, \beta} = \frac{-1}{n_2\pi_2(1-\pi_2)}. \end{aligned}$$

Note that the estimates are correlated even though the underlying likelihood is the product of two independent likelihoods. That is because each of these two binomial likelihoods involves the intercept α .

Also note that $\hat{\beta} = \log(\hat{O}\hat{R})$ and $V(\hat{\beta})$ is the same as Woolf's expression for the variance of the estimated log odds ratio obtained in Section 2.3.3 using the δ -method. Expressing the probabilities using the logistic function, the estimated variance is

$$\hat{V}(\hat{\beta}) = \frac{\left(1 + e^{\hat{\alpha} + \hat{\beta}}\right)^2}{n_1 e^{\hat{\alpha} + \hat{\beta}}} + \frac{\left(1 + e^{\hat{\alpha}}\right)^2}{n_2 e^{\hat{\alpha}}}. \quad (6.37)$$

Fig. 6.1 Two views of the likelihood surface in terms of the logit model parameters (α, β) for the data from the acid-dependent ulcer stratum of Example 4.1.



Substituting $e^{\hat{\alpha}} = b/d$, $e^{\hat{\beta}} = ad/bc$, and $e^{\hat{\alpha}+\hat{\beta}} = a/c$, it is readily shown that $\hat{V}(\hat{\beta})$ equals Woolf's estimated variance of the log odds ratio. Thus, large sample asymmetric confidence limits based on maximum likelihood estimation using this logit model are equivalent to those using log odds ratio in Section 2.3.3.

In Problem 6.4 we also show that the *MLE* of the log relative risk can be obtained from a model using a log link (function) in place of the logit in (6.16), with results equivalent to those presented in Section 2.3.2.

Example 6.1 *Ulcer Clinical Trial*

Consider only the first stratum in the ulcer clinical trial data from Example 4.1 with $a = 16$, $n_1 = 42$, $m_1 = 36$, and $N = 89$. Figure 6.1 displays the two views of the likelihood surface as a function of the parameters of the logit model parameters for these data. The values of the parameters at the peak of the surface are the *MLEs* $\hat{\alpha} = -0.3001$ and $\hat{\beta} = -0.1854$, the latter being the log odds ratio presented in Table 4.1. The elements of the estimated information are $\mathbf{I}(\boldsymbol{\theta})_{\hat{\alpha}} = 21.394$, and $\mathbf{I}(\boldsymbol{\theta})_{\hat{\beta}} = \mathbf{I}(\boldsymbol{\theta})_{\hat{\alpha},\hat{\beta}} = 9.9048$. These yield standard errors of the estimates $S.E.(\hat{\alpha}) = 0.2950$ and $S.E.(\hat{\beta}) = 0.4336$.

6.2.3 Tests of Significance

As described in Section A.7, either a large sample Wald, likelihood ratio, or score test can be applied in conjunction with maximum likelihood estimation. In the logit model for the 2×2 table, we wish to test the null hypothesis $H_0: \beta = 0$ against the

alternative $H_1: \beta \neq 0$. Since $\beta = 0 \Leftrightarrow \pi_1 = \pi_2$, this is equivalent to the hypothesis of no association in the 2×2 table employed previously.

6.2.3.1 Wald Test The Wald test of the log odds ratio in the logit model is defined as

$$X_W^2 = \frac{\hat{\beta}^2}{\hat{V}(\hat{\beta})}. \quad (6.38)$$

From (6.36), the variance of the estimate, that is obtained without restriction on the value of the parameter, or $\hat{V}(\hat{\beta}|H_1)$, is Woolf's estimated variance. Therefore, the Wald test is

$$X_W^2 = \frac{[\log(ad/bc)]^2}{\left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right]}. \quad (6.39)$$

Since the Wald test uses the variance estimated under the alternative, $\hat{V}(\hat{\beta}|H_1)$, this test does not equal any of the common tests for a 2×2 table presented in Chapter 2.

6.2.3.2 Likelihood Ratio Test The likelihood ratio test is based on the ratio of log likelihoods estimated with and without the parameter to be tested, β in this case, included in the model. Thus, for a 2×2 table, under a logit model, the likelihood ratio test for $H_0: \beta = 0$ is provided by

$$X_{LR}^2 = 2 \log[L(\hat{\alpha}, \hat{\beta})] - 2 \log[L(\hat{\alpha}_{|\beta=0})]. \quad (6.40)$$

From (6.21) the null log likelihood is

$$\ell(\alpha_{|\beta=0}) = \log[L(\alpha_{|\beta=0})] = m_1 \alpha - N \log(1 + e^\alpha), \quad (6.41)$$

and the score equation is

$$U(\theta)_{\alpha|\beta=0} = m_1 - N \frac{e^\alpha}{1 + e^\alpha}, \quad (6.42)$$

which, when set equal to zero, yields the estimate

$$\hat{\alpha}_0 = \hat{\alpha}_{|\beta=0} = \log(m_1/m_2). \quad (6.43)$$

Thus, the likelihood ratio test statistic is

$$\begin{aligned} X_{LR}^2/2 &= m_1 \hat{\alpha} + a \hat{\beta} - n_1 \log(1 + e^{\hat{\alpha} + \hat{\beta}}) - n_2 \log(1 + e^{\hat{\alpha}}) \\ &\quad - m_1 \hat{\alpha}_{|\beta=0} + N \log(1 + e^{\hat{\alpha}_{|\beta=0}}). \end{aligned} \quad (6.44)$$

Substituting the maximum likelihood estimates under the full and the null model and simplifying then yields the expression presented earlier in (2.97); see Problem 6.2.13.

6.2.3.3 Efficient Score Test The efficient score test for β may also be derived. Since the model involves two parameters (α and β) the score test for β is a $C(\alpha)$ test, as described in Section A.7.3.

By definition, the score equation for the intercept is zero when evaluated at the *MLE* for α under the null hypothesis, or $U(\boldsymbol{\theta})_{\hat{\alpha}|\beta=0} = 0$. Then evaluating the score equation for β under the null hypothesis $H_0: \beta = \beta_0 = 0$, given the estimate $\hat{\alpha}_0 = \hat{\alpha}_{|\beta=0}$ of the nuisance parameter in (6.43) yields

$$U(\hat{\boldsymbol{\theta}}_0)_\beta = a - \frac{n_1 e^{\hat{\alpha}_0}}{1 + e^{\hat{\alpha}_0}} = a - n_1 m_1 / N = a - \hat{E}(a|H_0). \quad (6.45)$$

This score equation can be viewed as arising from a *profile likelihood* of the form $L(\hat{\alpha}_{|\beta=0}, \beta_0) = L(\hat{\boldsymbol{\theta}}_0)$. Thus, the restricted *MLE* of the parameter vector under the null hypothesis is $\hat{\boldsymbol{\theta}}_0 = (\hat{\alpha}_{|\beta=0} \ \beta_0)^T$.

To obtain the score test we also need $\mathbf{I}(\boldsymbol{\theta})^{-1}$ evaluated under $H_0: \beta = 0$. Under H_0 , $\pi_1 = \pi_2 = \pi$ and from (6.42) it follows that

$$\pi = \frac{e^\alpha}{1 + e^\alpha} \hat{\equiv} \frac{m_1}{N}. \quad (6.46)$$

Substituting $\pi_1 = \pi_2 = \pi$ into the expressions for $\mathbf{I}(\boldsymbol{\theta})$ in (6.30) and evaluating at the estimate $\hat{\pi}$, we obtain the estimated information

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_0) = \hat{\pi}(1 - \hat{\pi}) \begin{bmatrix} N & n_1 \\ n_1 & n_1 \end{bmatrix}, \quad (6.47)$$

and the inverse

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1} = \left[\frac{1}{\hat{\pi}(1 - \hat{\pi})} \right] \left[\frac{1}{N n_1 - n_1^2} \right] \begin{bmatrix} n_1 & -n_1 \\ -n_1 & N \end{bmatrix}. \quad (6.48)$$

As shown in Section A.7.3, the efficient score test for $H_0: \beta = 0$, given the estimate of the nuisance parameter $\hat{\alpha}_{|\beta=0}$ evaluated under this hypothesis, is of the form

$$\begin{aligned} X^2 &= \left[U(\hat{\boldsymbol{\theta}}_0)_\alpha \ U(\hat{\boldsymbol{\theta}}_0)_\beta \right] \mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1} \left[U(\hat{\boldsymbol{\theta}}_0)_\alpha \ U(\hat{\boldsymbol{\theta}}_0)_\beta \right]^T \\ &= \left[0 \ U(\hat{\boldsymbol{\theta}}_0)_\beta \right] \mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1} \left[0 \ U(\hat{\boldsymbol{\theta}}_0)_\beta \right]^T \\ &= U(\hat{\boldsymbol{\theta}}_0)_\beta' \left[\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1} \right]_\beta U(\hat{\boldsymbol{\theta}}_0)_\beta, \end{aligned} \quad (6.49)$$

where, by definition, $U(\hat{\boldsymbol{\theta}}_0)_\alpha = [U(\boldsymbol{\theta})_\alpha]_{|\hat{\alpha}|\beta=0} = 0$. From (6.48), the desired element of the inverse estimated information corresponding to β evaluated under H_0 is

$$\left[\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1} \right]_\beta = \frac{N}{\hat{\pi}(1 - \hat{\pi}) n_1 n_2} = \frac{N^3}{m_1 m_2 n_1 n_2}. \quad (6.50)$$

The resulting efficient score test is

$$X^2 = \frac{[a - E(a)]^2}{\left(\frac{m_1 m_2 n_1 n_2}{N^3}\right)} = \frac{[a - \widehat{E}(a)]^2}{\widehat{V}_u(a)}, \quad (6.51)$$

that equals Cochran's test for the 2×2 table.

6.3 2×2 TABLE, CONDITIONALLY

Conditioning on both margins fixed, Section 2.4.2 shows that the likelihood function for the 2×2 table is the noncentral hypergeometric with parameter φ being the odds ratio. Again, it is convenient to parameterize the model in terms of the log odds ratio, $\beta = \log(\varphi)$. Then from (2.60), the log likelihood is

$$\ell(\beta) \propto a\beta - \log \left[\sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{n_2}{m_1 - i} e^{i\beta} \right] \quad (6.52)$$

with only one parameter β . The score equation then is

$$\begin{aligned} U(\beta) &= \frac{d\ell}{d\beta} = a - \frac{\sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{n_2}{m_1 - i} i e^{i\beta}}{\sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{n_2}{m_1 - i} e^{i\beta}} \\ &= a - E(a|\beta). \end{aligned} \quad (6.53)$$

The estimating equation for the *MLE* of β is $U(\beta) = 0$, which implies that

$$a \sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{n_2}{m_1 - i} e^{i\beta} = \sum_{i=a_\ell}^{a_u} \binom{n_1}{i} \binom{n_2}{m_1 - i} i e^{i\beta}, \quad (6.54)$$

for which there is no closed-form solution (see Birch, 1964).

However, the solution is readily obtained by an iterative procedure such as Newton-Raphson iteration described subsequently. To do so requires the expressions for the Hessian or observed information. Using a summary notation such that

$$C_i = \binom{n_1}{i} \binom{n_2}{m_1 - i}, \quad (6.55)$$

the observed information is

$$\begin{aligned} i(\beta) &= -\frac{d^2\ell}{d\beta^2} = \frac{[\sum_i C_i e^{i\beta}] [\sum_i i^2 C_i e^{i\beta}] - [\sum_i i C_i e^{i\beta}]^2}{[\sum_i C_i e^{i\beta}]^2} \\ &= \frac{\sum_i i^2 C_i e^{i\beta}}{\sum_i C_i e^{i\beta}} - \left[\frac{\sum_i i C_i e^{i\beta}}{\sum_i C_i e^{i\beta}} \right]^2. \end{aligned} \quad (6.56)$$

Thus, the observed and expected information is

$$i(\beta) = I(\beta) = E[a^2|\beta] - E(a|\beta)^2 = V(a|\beta). \quad (6.57)$$

Although the *MLE* must be solved for iteratively, the score test for $H_0: \beta = \beta_0 = 0$ is readily obtained directly. From (6.53), then

$$U(\beta_0) = a - E(a|H_0) = a - \frac{m_1 n_1}{N}. \quad (6.58)$$

When the information in (6.57) is evaluated under the null hypothesis with $\beta_0 = 0$, the expected information reduces to

$$I(\beta_0) = V(a|\beta_0) = V_c(a) = \frac{n_1 n_2 m_1 m_2}{N^2(N-1)} \quad (6.59)$$

based on the results presented in Section 2.6.3. Therefore, the score test is

$$X^2 = \frac{U(\beta_0)^2}{I(\beta_0)} = \frac{[a - E(a)]^2}{V_c(a)}, \quad (6.60)$$

that is the Mantel-Haenszel test.

6.4 SCORE-BASED ESTIMATE

In a meta-analysis of the effects of beta blockade on postmyocardial infarction mortality, Yusuf et al. (1985) and Peto (1987) employed a simple estimator of the log odds ratio that can be derived as a score equation-based estimator. Generalizations of this approach are also presented by Whitehead and Whitehead (1991), based on developments in the context of sequential analysis that are also described in Whitehead (1992).

In general, consider a likelihood in a single parameter, say θ , where the likelihood is parameterized such that the null hypothesis implies $H_0: \theta = \theta_0 = 0$. Then consider a Taylor's expansion of $U(\theta)$ about the null value θ_0 so that

$$U(\theta) = U(\theta_0) + (\theta - \theta_0) U'(\theta_0) + R_2(u) \quad (6.61)$$

for $u \in (\theta, \theta_0)$. Under a sequence of local alternatives where $\theta_n = \theta_0 + \delta/\sqrt{n}$, the remainder vanishes asymptotically. Since $\theta_0 = 0$, then asymptotically

$$U(\theta) \cong U(\theta_0) + \theta U'(\theta_0). \quad (6.62)$$

Taking expectations yields

$$E[U(\theta)] \cong E[U(\theta_0)] + \theta E[U'(\theta_0)]. \quad (6.63)$$

Since $E[U(\theta)] = 0$ and $E[U'(\theta_0)] = E[H(\theta_0)] = -I(\theta_0)$, then asymptotically

$$\theta \cong \frac{E[U(\theta_0)]}{I(\theta_0)}. \quad (6.64)$$

Therefore, the score-based estimate is

$$\hat{\theta} = \frac{U(\theta_0)}{I(\theta_0)}, \quad (6.65)$$

which is consistent for θ under local alternatives. As will be shown in Section 6.7, this score-based estimate equals the first step estimate in a Fisher Scoring iterative solution for the maximum likelihood estimate of the parameter when the initial (starting) estimate of the parameter is the value θ_0 .

From Section A.6.5, the asymptotic distribution of the score given the true value of θ is $U(\theta) \xrightarrow{d} \mathcal{N}[0, I(\theta)]$. From (6.62), asymptotically $U(\theta_0) \cong U(\theta) - \theta U'(\theta_0)$, where $U(\theta)$ converges in distribution to a normal variate, and from the law of large numbers, $-\theta U'(\theta_0) \xrightarrow{P} \theta I(\theta_0)$. Therefore, from Slutsky's theorem (A.43), asymptotically

$$U(\theta_0) \xrightarrow{d} \mathcal{N}[\theta I(\theta_0), I(\theta)], \quad (6.66)$$

and asymptotically

$$\hat{\theta} \xrightarrow{d} \mathcal{N}\left[\theta, \frac{I(\theta)}{I(\theta_0)^2}\right]. \quad (6.67)$$

Assuming that approximately $I(\theta) \doteq I(\theta_0)$, as will apply under local alternatives, then

$$\hat{\theta} \xrightarrow{d} \mathcal{N}\left[\theta, \frac{1}{I(\theta_0)}\right], \quad (6.68)$$

and asymptotically

$$V(\hat{\theta}) \cong I(\theta_0)^{-1}. \quad (6.69)$$

Example 6.2 Log Odds Ratio

Under the conditional hypergeometric distribution for the 2×2 table with fixed margins as in Section 6.3, the parameter of the likelihood is the odds ratio $\varphi = e^\beta$, where $\theta \equiv \beta = \log(OR)$. Using the log odds ratio rather than the odds ratio itself yields a value zero for the parameter under the null hypothesis ($\beta_0 = 0$) as required for the above. Therefore, the score-based estimate is

$$\hat{\beta} = \frac{U(\beta_0)}{I(\beta_0)}, \quad (6.70)$$

where the numerator is presented in (6.58) and the denominator in (6.59). Thus, the score-based estimate of the log odds ratio is

$$\log(\widehat{OR}) = \hat{\beta} = \frac{a - E(a|H_0)}{V_c(a)}, \quad (6.71)$$

with estimated variance

$$V(\hat{\beta}) \approx I(\beta_0)^{-1} = \frac{1}{V_c(a)} = \frac{N^2(N-1)}{n_1 n_2 m_1 m_2}. \quad (6.72)$$

Note that this variance estimate differs from Woolf's estimate that is also an inverse information-based estimate, as shown in Section 6.2.2.

Example 6.3 Ulcer Clinical Trial

Again, consider only the first stratum in the ulcer clinical trial data from Example 4.1. Then $\widehat{OR} = 0.8307$ and $\log \widehat{OR} = -0.18540$ with $V[\log \widehat{OR}] = 0.1880$ based on the δ -method. The score-based estimate of $\theta = \log(OR)$ is based on $a - E(a|H_0) = -0.98876$ and $V_c(a) = 5.4033$. The ratio yields $\widehat{\theta} = -0.18299$ with large sample variance $V(\widehat{\theta}) = 1/5.4033 = 0.18507$. Both quantities are close to those obtained previously using the δ -method.

6.5 STRATIFIED SCORE ANALYSIS OF INDEPENDENT 2×2 TABLES

Now consider the case of a stratified analysis with K independent 2×2 tables. Day and Byar (1979) used the logit model with a product binomial likelihood for each of the K tables to show that Cochran's test of association can be derived as a $C(\alpha)$ test. However, it is somewhat simpler to show that the Mantel-Haenszel test can be obtained using the conditional hypergeometric likelihood. Here both developments are described. As a problem, it is also shown that the Day and Byar approach can be employed to yield an efficient score test under a log risk model that equals the Radhakrishna test for a common relative risk presented in Section 4.7.2.

6.5.1 Conditional Mantel-Haenszel Test and the Score Estimate

In a stratified analysis with K independent 2×2 tables, assuming a common odds ratio under H_0 : $\beta_1 = \dots = \beta_K = \beta$, the conditional hypergeometric likelihood in (2.60) involves only a single parameter $\varphi = OR = e^\beta$. Then the resulting likelihood is the product of the hypergeometric likelihoods for each of the K strata

$$L(\beta) = \prod_{k=1}^K P(a_k | n_{1k}, m_{1k}, N_k, \beta). \quad (6.73)$$

Thus, the total score is $U(\beta) = \sum_k U_k(\beta)$ and the total information is $I(\beta) = \sum_k I_k(\beta)$. To obtain the MLE requires an iterative solution; see Example 6.7 in Section 6.5. However, the estimate is readily obtained under the null hypothesis.

Under H_0 : $\beta = \beta_0 = 0$, from (6.58), then

$$U(\beta_0) = \sum_k U_k(\beta_0) = \sum_k [a_k - E(a_k|H_0)] \quad (6.74)$$

and from (6.59)

$$I(\beta_0) = \sum_k I_k(\beta_0) = \sum_k V_c(a_k). \quad (6.75)$$

Therefore, the efficient score test is

$$X_S^2 = \frac{U(\beta_0)^2}{I(\beta_0)} = \frac{(\sum_k [a_k - E(a_k|H_0)])^2}{\sum_k V_c(a_k)}, \quad (6.76)$$

which equals the Mantel-Haenszel test $X_{C(MH)}^2$ in (4.8). Therefore, the Mantel-Haenszel test is an asymptotically efficient test of H_0 against $H_1: \beta_1 = \dots = \beta_K = \beta \neq 0$.

Also, the score-based estimate of the common log odds ratio $\beta = \log(OR)$ is

$$\hat{\beta} = \frac{U(\beta_0)}{I(\beta_0)} = \frac{\sum_k [a_k - E(a_k|H_0)]}{\sum_k V_c(a_k)}, \quad (6.77)$$

with estimated variance

$$\hat{V}(\hat{\beta}) = I(\beta_0)^{-1} = \frac{1}{\sum_k V_c(a_k)}. \quad (6.78)$$

Example 6.4 Ulcer Clinical Trial (continued)

For a stratified-adjusted analysis of the ulcer clinical trial data from Exercise 4.1 combined over strata, as shown in Table 4.3, the *MVLE* or logit estimate of the common log odds ratio for these three 2×2 tables is $\hat{\theta} = 0.493$, with estimated variance $\hat{V}(\hat{\theta}) = 0.0854$. The corresponding Mantel-Haenszel estimate of the log odds ratio is $\hat{\theta} = 0.491$, with the Robins et al. (1986) estimated variance $\hat{V}(\hat{\theta}) = 0.0813$. The score-based estimate involves $\sum_k (O_k - E_k) = 6.094$ and $\sum_k V_c(a_k) = 12.36$. These yield $\hat{\theta} = 0.493$, with variance $\hat{V}(\hat{\theta}) = 0.0809$, again comparable to the *MVLE* and Mantel-Haenszel estimates.

6.5.2 Unconditional Cochran Test as a $C(\alpha)$ Test

In many applications, the model involves one parameter of primary interest and one or more nuisance parameters, as illustrated by the efficient score test for the 2×2 table under a logit model in Section 6.2.3.3. In such cases, the score test for the parameter of interest is a $C(\alpha)$ test as described in Section A.7.3. Another such instance is the score test for the common odds ratio using a product binomial likelihood within each stratum described by Day and Byar (1979).

Consider a logit model for the K independent 2×2 tables with a common odds ratio $OR = e^\beta$, where the probabilities $\{\pi_{2k}\}$ are allowed to vary over strata. Applying a logit model to each table as in Section 6.2.2, the model yields a vector of $K + 1$ parameters $\boldsymbol{\theta} = (\alpha_1 \dots \alpha_K \beta)^T$, where α_k refers to the intercept or log background risk for the k th table ($k = 1, \dots, K$) and β represents the assumed common log odds ratio. Thus, the $\{\alpha_k\}$ are nuisance parameters that must be jointly estimated under the null hypothesis $H_0: \beta = \beta_0 = 0$. The resulting score test for β is a $C(\alpha)$ test and it equals the Cochran test for a common odds ratio presented in Section 4.2.4. The details are derived as a problem to illustrate this approach.

For the K independent 2×2 tables from each of the K independent strata, the compound product binomial likelihood is the product of the stratum-specific likelihoods, such as

$$L(\alpha_1, \dots, \alpha_K, \beta) = \prod_{k=1}^K L_k(\alpha_k, \beta). \quad (6.79)$$

For the k th stratum, the likelihood is a product-binomial

$$L_k(\alpha_k, \beta) = \binom{n_{1k}}{a_k} \binom{n_{2k}}{b_k} \pi_{1k}^{a_k} (1 - \pi_{1k})^{c_k} \pi_{2k}^{b_k} (1 - \pi_{2k})^{d_k}, \quad (6.80)$$

with probabilities

$$\pi_{1k} = \frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} \quad \text{and} \quad \pi_{2k} = \frac{e^{\alpha_k}}{1 + e^{\alpha_k}}, \quad (6.81)$$

where it is assumed that $\beta_1 = \dots = \beta_K = \beta$. In this case, since θ is a $K + 1$ vector, then a score test for $H_0: \beta = \beta_0 = 0$ entails a quadratic form evaluated as

$$X^2 = \mathbf{U}(\hat{\theta}_0)' \mathbf{I}(\hat{\theta}_0)^{-1} \mathbf{U}(\hat{\theta}_0), \quad (6.82)$$

where $\mathbf{U}(\hat{\theta}_0)$ is the score vector for the $K + 1$ parameters evaluated at the vector of estimates obtained under H_0 designated as $\hat{\theta}_0$. Thus, $\hat{\theta}_0 = (\hat{\alpha}_{1(0)} \dots \hat{\alpha}_{K(0)} 0)^T$, where $\hat{\alpha}_{k(0)}$ is the MLE for α_k that is obtained as the solution to $\frac{\partial \ell}{\partial \alpha_k}|_{\beta=\beta_0} = \frac{\partial \ell_k}{\partial \alpha_k}|_{\beta=\beta_0} = 0$.

However, when evaluated at the MLEs $\{\hat{\alpha}_{k(0)}\}$ estimated under the null hypothesis where $\beta = \beta_0$, then each respective element of the score vector is zero. Thus, the score vector

$$\mathbf{U}(\hat{\theta}_0) = \left[U(\hat{\theta}_0)_{\alpha_1} \dots U(\hat{\theta}_0)_{\alpha_K} U(\hat{\theta}_0)_{\beta} \right] = \left[0 \dots 0 U(\hat{\theta}_0)_{\beta} \right] \quad (6.83)$$

is a single random variable augmented by zeros, where

$$U(\hat{\theta}_0)_{\beta} = \left[U(\theta)_{\beta} \right] |_{\hat{\alpha}, \beta=\beta_0} = \frac{\partial \ell}{\partial \beta} |_{\hat{\alpha}, \beta=\beta_0}, \quad (6.84)$$

and $\hat{\alpha} = (\hat{\alpha}_{1(0)} \dots \hat{\alpha}_{K(0)})^T$. Then the quadratic form reduces to simply

$$X^2 = \left[\mathbf{U}(\hat{\theta}_0)_{\beta} \right]^2 \mathbf{I}(\hat{\theta}_0)^{\beta}, \quad (6.85)$$

where $\mathbf{I}(\hat{\theta}_0)^{\beta} = [\mathbf{I}(\hat{\theta}_0)^{-1}]_{\beta}$ is the diagonal element of the inverse information matrix corresponding to the parameter β and evaluated at $\beta = \beta_0$. Thus, X^2 is distributed asymptotically as chi-square on 1 df.

As a problem it is readily shown that $\hat{\alpha}_{k(0)} = \log(m_{1k}/m_{2k})$ and that the resulting $C(\alpha)$ score statistic for β is

$$U(\hat{\theta}_0)_{\beta} = \sum_k \left(a_k - \frac{m_{1k} n_{1k}}{N_k} \right). \quad (6.86)$$

Then under H_0 : $\pi_1 = \pi_2 = \pi$, from (6.30) the information matrix has elements

$$\begin{aligned}\mathbf{I}(\boldsymbol{\theta}_0)_{\alpha_k} &= \pi_k (1 - \pi_k) N_k \hat{=} \frac{m_{1k} m_{2k}}{N_k} \\ \mathbf{I}(\boldsymbol{\theta}_0)_{\alpha_k, \beta} &= \pi_k (1 - \pi_k) n_{1k} \hat{=} \frac{m_{1k} m_{2k} n_{1k}}{N_k^2} \\ \mathbf{I}(\boldsymbol{\theta}_0)_{\beta} &= \sum_k \pi_k (1 - \pi_k) n_{1k} \hat{=} \sum_k \frac{m_{1k} m_{2k} n_{1k}}{N_k^2}\end{aligned}\quad (6.87)$$

and $\mathbf{I}(\boldsymbol{\theta}_0)_{\alpha_k, \alpha_\ell} = 0$ for $k \neq \ell$ since the strata are independent. Therefore, the estimated information matrix is a patterned matrix of the form

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_0) = \begin{bmatrix} \mathbf{I}(\hat{\boldsymbol{\theta}}_0)_\alpha & \mathbf{I}(\hat{\boldsymbol{\theta}}_0)_{\alpha, \beta} \\ \mathbf{I}(\hat{\boldsymbol{\theta}}_0)_{\alpha, \beta}^T & \mathbf{I}(\hat{\boldsymbol{\theta}}_0)_\beta \end{bmatrix}, \quad (6.88)$$

where $\mathbf{I}(\hat{\boldsymbol{\theta}}_0)_\alpha = \text{diag}\{\mathbf{I}(\hat{\boldsymbol{\theta}}_0)_{\alpha_1}, \dots, \mathbf{I}(\hat{\boldsymbol{\theta}}_0)_{\alpha_K}\}$ is a $K \times K$ diagonal matrix.

The required element of the inverse matrix $\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^\beta = [\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1}]_\beta$ is obtained using the expression for the inverse of a patterned matrix presented in (A.3) of Section A.1.2. Then it is readily shown that

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^\beta = \frac{1}{\sum_k \hat{V}_u(a_k)}, \quad (6.89)$$

and that

$$X^2 = \frac{\left(\sum_k [a_k - E(a_k|H_0)]\right)^2}{\sum_k \hat{V}_u(a_k)}, \quad (6.90)$$

that is Cochran's test $X_{U(C)}^2$ in (4.12). Note that in order to compute the test one does not need to solve for the estimates of the nuisance parameters, the $\{\hat{\alpha}_k\}$.

Gart (1985) used a similar model parameterized using the log link, where $\pi_{1k} = e^{\alpha_k + \beta}$ and $\pi_{2k} = e^{\alpha_k}$, in which case $\beta = \log(RR)$. The resulting score test equals Radhakrishna's test for a common relative risk (see Problem 6.7). Gart and Tarone (1983) present general expressions for score tests for the exponential family of likelihoods, which includes the binomial with a logit link as above but not the relative risk model of Gart (1985).

6.6 MATCHED PAIRS

6.6.1 Unconditional Logit Model

Now consider a likelihood-based analysis of a sample of N pairs of observations that have been matched with respect to the values of a set of covariates, or that are intrinsically matched such as before-and-after measurements from the same subject. As in Sections 5.3 and 5.4, consider the analysis of a prospective study in which pairs of exposed and nonexposed individuals are compared with respect to the risk

of developing a disease, or experiencing a positive outcome. Let the binary indicator variable X denote the conditions to be compared. For the j th member ($j = 1, 2$) of the i th pair ($i = 1, \dots, N$), $x_{ij} = 1$ denotes the exposed (E) member and $x_{ij} = 0$ denotes nonexposed (\bar{E}) member. Then let Y denote the outcome of each pair member, where $y_{ij} = 1$ if the ij th subject has a positive outcome (D), 0 if not (\bar{D}). To simplify, we use the convention that $j = 1$ refers to the exposed (E) member and $j = 2$ to the nonexposed (\bar{E}) member; i.e., $x_{i1} = 1$ and $x_{i2} = 0$ for all pairs. Then the data for the i th pair form a 2×2 table, and when aggregated over all N pairs yields

		\bar{E} ($x_{i2} = 0$)	
		D ($y_{i2} = 1$)	\bar{D} ($y_{i2} = 0$)
E ($x_{i1} = 1$)	D ($y_{i1} = 1$)	$e = n_{11}$	$f = n_{12}$
\bar{D} ($y_{i1} = 0$)	\bar{D} ($y_{i1} = 1$)	$g = n_{21}$	$h = n_{22}$

(6.91)

where the underlying probabilities are $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ for the four cells of the table for all pairs. The number of discordant pairs is $M = n_{12} + n_{21} = f + g$, where $E(M) = \pi_d N$, $\pi_d = \pi_{12} + \pi_{21}$ being the probability of a discordant pair.

For the i th pair a simple logistic model can be used to describe the odds ratio, allowing for the underlying probabilities of the outcome to vary over pairs. Within the i th pair, since the exposed and nonexposed individuals were sampled independently given the matching covariate values, then as in Section 5.4.1 we can define the conditional probabilities of the outcome as

$$\begin{aligned}\pi_{i1} &= P(y_{i1} = 1 | x_{i1} = 1) = P_i(D|E) = P(D|E, z_i) \\ \pi_{i2} &= P(y_{i2} = 1 | x_{i2} = 0) = P_i(D|\bar{E}) = P(D|\bar{E}, z_i).\end{aligned}\quad (6.92)$$

In terms of the notation of Section 5.4.1, for a matched pair with shared covariate value z , then $\pi_{i1} \equiv \pi_{1\bullet|z}$ and $\pi_{i2} \equiv \pi_{\bullet1|z}$.

Now, as originally suggested by Cox (1958b), we can adopt a logistic model assuming a constant odds ratio for all N pairs of the form

$$\log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \alpha_i + \beta x_{ij}, \quad i = 1, \dots, N; j = 1, 2. \quad (6.93)$$

Therefore, the log odds ratio, or difference in logits for the exposed and nonexposed members, is

$$\log \left(\frac{\pi_{i1}}{1 - \pi_{i1}} \right) - \log \left(\frac{\pi_{i2}}{1 - \pi_{i2}} \right) = \alpha_i + \beta x_{i1} - (\alpha_i + \beta x_{i2}) = \beta. \quad (6.94)$$

As illustrated by the Mantel-Haenszel construction for matched pairs in Section 5.4.4, within each pair the members are conditionally independent so that the likelihood is a product binomial for the i th independent pair. Using the logistic function

(inverse logits) for each probability, the likelihood is

$$L_i(\alpha_i, \beta) \propto \left(\frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \right)^{y_{i1}} \left(\frac{1}{1 + e^{\alpha_i + \beta}} \right)^{1-y_{i1}} \left(\frac{e^{\alpha_i}}{1 + e^{\alpha_i}} \right)^{y_{i2}} \left(\frac{1}{1 + e^{\alpha_i}} \right)^{1-y_{i2}}. \quad (6.95)$$

Therefore, the total log likelihood for the complete sample of N matched pairs up to a constant is $\ell = \sum_i \log L_i = \sum_i \ell_i$, which equals

$$\ell = \sum_i \alpha_i (y_{i1} + y_{i2}) + \sum_i y_{i1} \beta - \sum_i \log(1 + e^{\alpha_i}) - \sum_i \log(1 + e^{\alpha_i + \beta}). \quad (6.96)$$

The resulting score equations for the parameters $\theta = (\alpha_1 \dots \alpha_N \ \beta)^T$ are

$$U(\theta)_{\alpha_i} = \frac{\partial \ell}{\partial \alpha_i} = \frac{\partial \ell_i}{\partial \alpha_i} = (y_{i1} + y_{i2}) - \frac{e^{\alpha_i}}{1 + e^{\alpha_i}} - \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}}, \quad (6.97)$$

for $i = 1, \dots, N$, and

$$U(\theta)_{\beta} = \frac{\partial \ell}{\partial \beta} = \sum_i y_{i1} - \sum_i \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}}. \quad (6.98)$$

Thus, in order to estimate the log odds ratio β , it is also necessary that we jointly estimate the N nuisance parameters $\{\alpha_i\}$ as well. Since the number of parameters equals $N + 1$, this yields an unstable estimate of β , the bias of which increases with N (cf. Cox and Hinkley, 1974).

6.6.2 Conditional Logit Model

However, Cox (1958b) showed that the assumed common log odds ratio β can be estimated without simultaneously estimating the N nuisance parameters $\{\alpha_i\}$ by applying the principle of conditioning originally attributed to Fisher (cf. Fisher, 1956). Since the maximum likelihood estimator for a parameter is a function of the asymptotically *sufficient statistic* for that parameter, then from the log likelihood in (6.96), the number of positive responses within the i th pair, $S_i = y_{i1} + y_{i2}$, is the sufficient statistic for the parameter α_i , where $S_i = 0, 1$ or 2 . In a loose sense this means that the ancillary statistic S_i captures all the information about the nuisance parameter α_i in the data and that inferences about the additional parameters in the data (the β) may be based on the conditional likelihood, after conditioning on the values of the ancillary parameter sufficient statistics.

The conditional likelihood then is $L(\beta)|_S = \prod_{i=1}^N L(\beta)_{i|S_i}$, where the conditional likelihood for the i th pair is

$$L(\beta)_{i|S_i} = P(y_{i1}, y_{i2} | S_i) = \frac{P(y_{i1}, y_{i2}, S_i)}{P(S_i)} = \frac{P(y_{i1}, y_{i2})}{P(S_i)}, \quad (6.99)$$

since S_i depends explicitly on the values of (y_{i1}, y_{i2}) . However, referring to the 2×2 table in (6.91), $y_{i1} = y_{i2} = 0$ for the h pairs where $S_i = 0$. Thus, for these pairs, the

conditional likelihood is $P(y_{i1} = 0, y_{i2} = 0 \mid S_i = 0) = 1$. Likewise, $y_{i1} = y_{i2} = 1$ for the e pairs where $S_i = 2$, so that $P(y_{i1} = 1, y_{i2} = 1 \mid S = 2) = 1$. Therefore, each concordant pair ($S_i = 0$ or 2) contributes a constant (unity) to the conditional likelihood and provides no information about β .

Conversely, there are M discordant pairs where $S_i = 1$, f of which with probability

$$P(y_{i1} = 1, y_{i2} = 0) = \pi_{i1}(1 - \pi_{i2}) = \left(\frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \right) \left(\frac{1}{1 + e^{\alpha_i}} \right), \quad (6.100)$$

and g with probability

$$P(y_{i1} = 0, y_{i2} = 1) = (1 - \pi_{i1})\pi_{i2} = \left(\frac{1}{1 + e^{\alpha_i + \beta}} \right) \left(\frac{e^{\alpha_i}}{1 + e^{\alpha_i}} \right). \quad (6.101)$$

The total probability of a discordant pair, or $P(S_i = 1)$, is the sum of these two probabilities. Therefore, the conditional likelihood is

$$L_{|S} = \prod_{i=1}^N \frac{P(y_{i1}, y_{i2})}{P(S_i)} = \prod_{i:S_i=1} \frac{P(y_{i1}, y_{i2})}{P(S_i = 1)} = L_{|S=1}, \quad (6.102)$$

since only the discordant pairs contribute to the likelihood.

The resulting conditional likelihood can be expressed as

$$L(\beta)_{|S=1} = \prod_{i:S_i=1} \left[\frac{P(y_{i1} = 1, y_{i2} = 0)}{P(S_i = 1)} \right]^{y_{i1}(1-y_{i2})} \left[\frac{P(y_{i1} = 0, y_{i2} = 1)}{P(S_i = 1)} \right]^{y_{i2}(1-y_{i1})} \quad (6.103)$$

where one, but not both, of the exponents equals 1, the other 0. Then, for the f pairs where $y_{i1}(1 - y_{i2}) = 1$, as a problem it is readily shown that

$$\frac{P(y_{i1} = 1, y_{i2} = 0)}{P(S_i = 1)} = \frac{\pi_{i1}(1 - \pi_{i2})}{\pi_{i1}(1 - \pi_{i2}) + \pi_{i2}(1 - \pi_{i1})} = \frac{e^\beta}{1 + e^\beta} \quad (6.104)$$

and for the g pairs where $y_{i2}(1 - y_{i1}) = 1$ that

$$\frac{P(y_{i2} = 1, y_{i1} = 0)}{P(S_i = 1)} = \frac{\pi_{i2}(1 - \pi_{i1})}{\pi_{i1}(1 - \pi_{i2}) + \pi_{i2}(1 - \pi_{i1})} = \frac{1}{1 + e^\beta}. \quad (6.105)$$

Thus, the conditional likelihood is

$$\begin{aligned} L_c(\beta) = L(\beta)_{|S=1} &= \prod_{i:S_i=1} \left[\frac{e^\beta}{1 + e^\beta} \right]^{y_{i1}(1-y_{i2})} \left[\frac{1}{1 + e^\beta} \right]^{y_{i2}(1-y_{i1})} \\ &= \left(\frac{e^\beta}{1 + e^\beta} \right)^f \left(\frac{1}{1 + e^\beta} \right)^g, \end{aligned} \quad (6.106)$$

which depends only on β and not on the nuisance parameters $\{\alpha_i\}$ that appear in the unconditional likelihood.

The resulting score equation for β is

$$U(\beta) = \frac{d \log L_c(\beta)}{d\beta} = \frac{d\ell_c}{d\beta} = f - \frac{Me^\beta}{1 + e^\beta}, \quad (6.107)$$

which, when set to zero, yields the *MLE*

$$\hat{\beta} = \log \left(\frac{f}{g} \right). \quad (6.108)$$

The observed information is

$$i(\beta) = \frac{-d^2 \log L_c(\beta)}{d\beta^2} = M \left[\frac{e^\beta}{(1 + e^\beta)^2} \right], \quad (6.109)$$

and the expected information is

$$I(\beta) = \frac{E(M) e^\beta}{(1 + e^\beta)^2}. \quad (6.110)$$

Evaluating the expected information at the *MLE* and conditioning on the observed value of M yields the estimated information

$$I(\hat{\beta}) = \frac{fg}{M}. \quad (6.111)$$

Thus, the large sample variance of the estimate is

$$V(\hat{\beta}) = I(\hat{\beta})^{-1} = \frac{M}{fg}, \quad (6.112)$$

as obtained in Section 5.4.

It is also instructive to note that the conditional likelihood in (6.106) is identical to the marginal conditional binomial likelihood obtained by conditioning on the total number of discordant pairs (M) as presented in (5.30). That conditional likelihood is

$$L(\beta)_{|M} = \binom{M}{f} \left(\frac{\pi_{12}}{\pi_d} \right)^f \left(\frac{\pi_{21}}{\pi_d} \right)^g, \quad (6.113)$$

which, apart from the constant, equals the conditional logistic regression model likelihood, where

$$\frac{\pi_{12}}{\pi_d} = \frac{e^\beta}{1 + e^\beta} \quad \text{and} \quad \frac{\pi_{21}}{\pi_d} = \frac{1}{1 + e^\beta}. \quad (6.114)$$

6.6.3 Conditional Likelihood Ratio Test

The conditional likelihood in (6.106) then provides a likelihood ratio test for the hypothesis $H_0: \beta = 0$ of no association between exposure and the outcome in the population after matching on the matching covariates. To simplify notation, designate the conditional likelihood in (6.106) as simply $L(\beta)$. Under this conditional logit model, the null conditional likelihood has no intercept, and thus $\log(L)$ is simply $\ell(\beta)|_{\beta=0} = -M \log(2)$. The conditional likelihood evaluated at the conditional MLE $\hat{\beta}$ then yields the likelihood ratio test statistic

$$X_{LR}^2 = -2 \log \left[\frac{(M/2)^M}{fg} \right], \quad (6.115)$$

which is asymptotically distributed as chi-square on 1 *df*.

6.6.4 Conditional Score Test

Now consider the score test for $H_0: \beta = 0$ in the conditional logit model. It is readily shown that

$$U(\beta)|_{\beta=0} = (f - g)/2, \quad (6.116)$$

and that

$$I(\beta)|_{\beta=0} = \frac{E(M)}{4} \hat{=} \frac{M}{4}. \quad (6.117)$$

Therefore, the efficient score test is

$$X^2 = \frac{(f - g)^2}{M}, \quad (6.118)$$

that is McNemar's test.

Example 6.5 Hypothetical Data

Consider the simple example where the matched 2×2 table has discordant frequencies $f = 8$ and $g = 15$. Then the likelihood ratio test from (6.115) yields $X_{LR}^2 = 2.16$ and the score or McNemar's test yields $X^2 = 2.13$, neither of which is significant at the 0.05 level.

6.6.5 Matched Case-Control Study

Now consider a case-control study with matched pairs consisting of a case with the disease (D) and a control without the disease (\bar{D}), where the prior exposure status of each has been determined retrospectively (E or \bar{E}). In this model, X is the binary independent variable, which denotes being a case (D) versus a control (\bar{D}); and Y is the binary dependent variable which denotes having been exposed (E) versus not exposed (\bar{E}). Then using the same construction as in Section 6.6.1, for the j th member of the i th pair, the conditional probability of exposure is

$$\phi_{ij} = P(E|x_{ij}) = P(y_{ij} = 1|x_{ij}) \quad (6.119)$$

for $i = 1, \dots, N$ and $j = 1, 2$. Note that the $\{x_{ij}\}$ and the $\{y_{ij}\}$ refer to the independent and dependent variables, respectively, under retrospective sampling.

As for the analysis of matched pairs from a prospective study in Section 6.6.2, assume a logit model for these conditional probabilities such that

$$\log \left(\frac{\phi_{ij}}{1 - \phi_{ij}} \right) = \alpha_i + \beta x_{ij}. \quad (6.120)$$

Then the marginal retrospective odds ratio for the i th matched pair is

$$OR_{retro} = \frac{\phi_{i1} (1 - \phi_{i2})}{\phi_{i2} (1 - \phi_{i1})} = \frac{P_i(E|D) P_i(\bar{E}|\bar{D})}{P_i(\bar{E}|D) P_i(E|\bar{D})} = e^\beta. \quad (6.121)$$

Using the prior probabilities or prevalences $P(D)$ and $P(\bar{D})$ and applying Bayes' theorem, as in Section 5.5, it is readily shown that the retrospective odds ratio equals the prospective odds ratio:

$$OR_{retro} = OR = \frac{\pi_{i1} (1 - \pi_{i2})}{\pi_{i2} (1 - \pi_{i1})} = e^\beta. \quad (6.122)$$

Therefore, the results in Sections 6.6.1 to 6.6.4 also apply to the analysis of odds ratios in a matched case-control study. These conclusions are identical to those described in Section 5.5, where it is shown that $\hat{\beta} = \log(\widehat{OR}) = \log(f/g)$. An example is presented in Example 5.8.

6.7 ITERATIVE MAXIMUM LIKELIHOOD

In many cases, a model is formulated in terms of a set of parameters $\boldsymbol{\theta} = (\theta_1 \dots \theta_p)$, where the *MLE* estimating equation for $\boldsymbol{\theta}$ cannot be solved in closed form. For example, under an appropriate model, the probabilities of a C category multinomial $\{\pi_i\}$ may be a function of a smaller set of $p \leq C$ parameters. Although the p estimating equations may admit a simultaneous unique solution, meaning that the likelihood is *identifiable* in the parameters, no direct solution may exist, or may be computed tractably. In such cases, an iterative or recursive computational method is required to arrive at the solution vector. The two principal approaches are Newton-Raphson iteration and Fisher scoring, among many. Readers are referred to Thisted (1988), among others, for a comprehensive review of the features of these and other methods.

6.7.1 Newton-Raphson (or Newton's Method)

Consider the scalar parameter case and assume that we have an initial guess as to the value of the parameter that is "in the neighborhood" of the desired solution $\hat{\theta}$. Taking a Taylor's expansion of the estimating equation $U(\hat{\theta}) = 0$ about the starting value $\hat{\theta}^{(0)}$, we have

$$0 = U(\hat{\theta}) = U(\hat{\theta}^{(0)}) + (\hat{\theta} - \hat{\theta}^{(0)}) U'(\hat{\theta}^{(0)}) + R_2, \quad (6.123)$$

which implies that

$$\hat{\theta} \cong \hat{\theta}^{(0)} - \frac{U(\hat{\theta}^{(0)})}{U'(\hat{\theta}^{(0)})} = \hat{\theta}^{(0)} + \frac{U(\hat{\theta}^{(0)})}{i(\hat{\theta}^{(0)})}. \quad (6.124)$$

A negative step size based on the Hessian $U'(\hat{\theta})$ is equivalent to a positive step size using the observed information $i(\hat{\theta})$. This equation is then applied iteratively until $\hat{\theta}$ converges to a constant (the desired solution) using the sequence of equations

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - \frac{U(\hat{\theta}^{(i)})}{U'(\hat{\theta}^{(i)})} = \hat{\theta}^{(i)} + \frac{U(\hat{\theta}^{(i)})}{i(\hat{\theta}^{(i)})}, \quad (6.125)$$

for $i = 0, 1, 2, \dots$

For the case where θ is a vector, the corresponding sequence of equations is characterized as

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - \left[U'(\hat{\theta}^{(i)}) \right]^{-1} U(\hat{\theta}^{(i)}) \quad (6.126)$$

$$= \hat{\theta}^{(i)} + i(\hat{\theta}^{(i)})^{-1} U(\hat{\theta}^{(i)}). \quad (6.127)$$

By definition, the *MLE* is obtained when the solution $U(\hat{\theta}^{(i+1)}) = 0$.

In the univariate case this can be pictured graphically, as shown in Figure 6.2, where $-U(\hat{\theta}^{(i)})/U'(\hat{\theta}^{(i)})$ is the step size and $-U(\hat{\theta}^{(i)})$ determines the direction of the step as shown in (6.124). The initial estimate is often the value expected under an appropriate null hypothesis or may be an estimate provided by a simple noniterative moment estimator when such exists. Then the above expression is equivalent to determining the tangent to the log likelihood contour and projecting it to the abscissa to determine the next iterative estimate. This process continues until the solution converges to a constant to the specified precision.

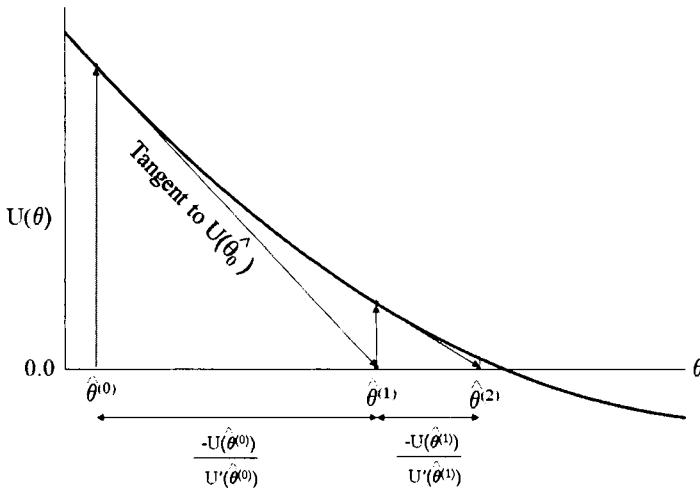
The Newton-Raphson iteration is quadratic convergent with respect to the rate at which $U(\hat{\theta}) \rightarrow 0$. However, it is also sensitive to the choice of the initial or starting value.

6.7.2 Fisher Scoring (Method of Scoring)

Fisher (1935) proposed a similar procedure that he termed the *method of scoring*. Fisher scoring employs the estimated expected information to determine the step size rather than the Hessian (or observed information). Here the updated iterative estimate at the $(i + 1)$ th step is obtained as

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} + \frac{U(\hat{\theta}^{(i)})}{I(\hat{\theta}^{(i)})}, \quad i = 0, 1, 2, \dots \quad (6.128)$$

Fig. 6.2 Two steps of a Newton-Raphson iterative solution of a score estimating equation.



In the multiparameter case the vector estimate is obtained as

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} + \left[I \left(\hat{\theta}^{(i)} \right) \right]^{-1} U \left(\hat{\theta}^{(i)} \right). \quad (6.129)$$

Again note that step size is determined by $I(\hat{\theta}^{(i)})$ and the direction by $U(\hat{\theta}^{(i)})$.

Scoring methods generally do not converge to a solution as quickly as does Newton-Raphson. However, the Newton-Raphson method is much more sensitive to the choice of initial values. If the initial estimate $\hat{\theta}^{(0)}$ is far off, it is more likely to provide a divergent solution.

Example 6.6 Recombination Fraction

These computations are illustrated by the example from Fisher (1925), described in Example 2.15, of a quadrinomial distribution where the probabilities of the four possible combinations of two binary traits are expressed as a function of a single parameter θ as given in (2.126). The recombination fraction is $\sqrt{\theta}$, $0 < \theta < 1$, (cf. Thisted, 1988). Under the null hypothesis of Mendelian inheritance, H_0 : $\theta = \theta_0 = 0.25$. In Fisher's example, $N = 3839$, $x_1 = 1997$, $x_2 = 906$, $x_3 = 904$, and $x_4 = 32$. Then the likelihood function is a quadrinomial with

$$L = \frac{N!}{x_1! x_2! x_3! x_4!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3} \pi_4^{x_4} \quad (6.130)$$

as a function of these frequencies and the probabilities described in (2.126). Therefore,

$$\ell \propto \sum_{j=1}^4 x_j \log(\pi_j) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta), \quad (6.131)$$

and the score equation is

$$U(\theta) = \frac{x_1}{2 + \theta} - \frac{x_2 + x_3}{1 - \theta} + \frac{x_4}{\theta}, \quad (6.132)$$

which is a quadratic function in θ . Although we could solve for θ directly as the positive root of the quadratic equation, we will proceed to solve for the *MLE* of θ using iterative computations.

The Hessian is

$$U'(\theta) = - \left[\frac{x_1}{(2 + \theta)^2} + \frac{x_2 + x_3}{(1 - \theta)^2} + \frac{x_4}{\theta^2} \right] = - \left[\frac{1997}{(2 + \theta)^2} + \frac{1810}{(1 - \theta)^2} + \frac{32}{\theta^2} \right] \quad (6.133)$$

and the expected information function is

$$\begin{aligned} I(\theta) &= -E[U'(\theta)] \quad (6.134) \\ &= N \left[\frac{\pi_1}{(2 + \theta)^2} + \frac{\pi_2 + \pi_3}{(1 - \theta)^2} + \frac{\pi_4}{\theta^2} \right] = \frac{N}{4} \left[\frac{1}{2 + \theta} + \frac{2}{1 - \theta} + \frac{1}{\theta} \right]. \end{aligned}$$

In many problems, the iterative solution can be accelerated by starting with a good initial noniterative estimate of the parameter, such as a moment estimator. For this model, a simple moment estimator for θ is readily obtained (see Problem 6.9). Each frequency provides a separate moment estimator of θ , for example

$$E(x_1) = \frac{N(2 + \theta)}{4} \Rightarrow \hat{\theta}_1 = \frac{4x_1}{N} - 2. \quad (6.135)$$

Averaging these estimators from the four frequencies provides the initial moment estimator

$$\hat{\theta}^{(0)} = (x_1 - x_2 - x_3 + x_4)/N \quad (6.136)$$

which yields $\hat{\theta}^{(0)} = 0.05704611$ for this example.

Using $\hat{\theta}^{(0)}$ at the initial step yields

$$U(\hat{\theta}^{(0)}) = \frac{1997}{2.057} - \frac{906 + 904}{1 - 0.057} + \frac{32}{0.057} = -387.74068038, \quad (6.137)$$

and

$$U'(\hat{\theta}^{(0)}) = - \left[\frac{1997}{(2.057)^2} + \frac{1810}{(1 - 0.057)^2} + \frac{32}{(0.057)^2} \right] = -12340.84. \quad (6.138)$$

Table 6.1 Newton-Raphson iterative solution for the *MLE*.

i	$\hat{\theta}^{(i)}$	$U(\hat{\theta}^{(i)})$	$U'(\hat{\theta}^{(i)})$
0	0.05704611	-387.740680	-12340.838
1	0.02562679	376.956469	-51119.245
2	0.03300085	80.1936782	-31802.060
:	:	:	:
6	0.03571230	-0.00000000	-27519.223

Therefore, the first step in Newton-Raphson iteration yields the revised estimate

$$\hat{\theta}^{(1)} = \hat{\theta}^{(0)} - \left(\frac{-387.74}{-12340.84} \right) = 0.0256268.$$

Table 6.1 presents the values of $\hat{\theta}^{(i)}$ at each step of the Newton-Raphson iteration, the values of the efficient score $U(\hat{\theta}^{(i)})$ and the values of the Hessian $U'(\hat{\theta}^{(i)})$. Newton-Raphson required six iterations to reach a solution to eight decimal places. The *MLE* is $\hat{\theta} = 0.03571230$ and the estimate of the observed information is $i(\hat{\theta}) = -U'(\hat{\theta}) = 27519.223$. Therefore, the estimated variance of the estimate obtained as the inverse of the estimated observed information is $\hat{V}_i(\hat{\theta}) = 1/27519.223 = 3.6338 \times 10^{-5}$.

Table 6.2 presents each step of the Fisher scoring iterative procedure with the estimated information that determines the step size. Fisher scoring required eight iterations. The initial estimated information is

$$I(\hat{\theta}^{(0)}) = \frac{3839}{4} \left[\frac{1}{2.057} + \frac{2}{1 - 0.057} + \frac{1}{0.057} \right] = 19326.3, \quad (6.139)$$

which yields the first-step estimate $\hat{\theta}^{(1)} = 0.057 + (-387.74/19326.3) = 0.03698$. The remaining iterative calculations are shown in Table 6.2. At the final convergent iteration the estimated expected information is $I(\hat{\theta}) = 29336.538$, which yields a slightly different estimate of the variance

$$\hat{V}_I(\hat{\theta}) = \frac{1}{29336.538} = 3.4087 \times 10^{-5}. \quad (6.140)$$

Now consider a test of independence of loci or $H_0: \theta = 0.25$. The Wald Test using the variance estimate $\hat{V}_I(\hat{\theta}) = I(\hat{\theta})$ obtained from the estimated information is

$$\begin{aligned} X^2 &= (\hat{\theta} - 0.25)^2 / \hat{V}_I(\hat{\theta}) = (\hat{\theta} - 0.25)^2 I(\hat{\theta}) \\ &= (0.0357 - 0.25)^2 (29336.54) = 1347.1, \end{aligned} \quad (6.141)$$

Table 6.2 Fisher scoring iterative solution for the *MLE*.

i	$\hat{\theta}^{(i)}$	$U(\hat{\theta}^{(i)})$	$I(\hat{\theta}^{(i)})$
0	0.05704611	-387.740680	19326.302
1	0.03698326	-33.882673	28415.310
2	0.03357908	-2.157202	29277.704
:	:	:	:
6	0.03571232	-0.00051375	29336.524
7	0.03571230	-0.00003182	29336.537
8	0.03571230	-0.00000197	29336.538

whereas the efficient score test is

$$X^2 = U(0.25)^2 / I(0.25) = (-1397.78)^2 / 6824.89 = 286.27,$$

both of which are highly significant on 1 *df*. Even though the score test is less significant, it is preferred because it uses the variance estimated under the null rather than under the alternative hypothesis as in the Wald test. Thus, the size (Type I error probability) of the Wald test may be inflated.

The likelihood ratio test is computed as $X^2 = -2[\log L(0.25) - \log L(\hat{\theta})]$, where the log likelihood is $\log L(\theta) \propto \sum_j x_j \log[\pi_j(\theta)] = \tilde{\ell}(\theta)$. Evaluating each yields $\tilde{\ell}(0.25) = -4267.62$ under the tested hypothesis and $\tilde{\ell}(\hat{\theta}) = \tilde{\ell}(0.0357) = -4074.88$ globally. Therefore, the likelihood ratio test is $X^2 = 2(4267.62 - 4074.88) = 385.48$.

Example 6.7 Conditional MLE, Ulcer Clinical Trial

An iterative procedure such as Newton-Raphson is required to compute the *MLE* of the odds ratio from the conditional hypergeometric distribution for a 2×2 table with fixed margins, or for a set of independent tables in a stratified model with a common odds ratio. From Section 6.5.1, the score vector for K independent strata is

$$U(\beta) = \sum_{k=1}^K \left[a_k - \frac{\sum_{i=a_{\ell k}}^{a_{u k}} \binom{n_{1k}}{i} \binom{n_{2k}}{m_{1k} - i} i e^{i\beta}}{\sum_{i=a_{\ell k}}^{a_{u k}} \binom{n_{1k}}{i} \binom{n_{2k}}{m_{1k} - i} e^{i\beta}} \right] = \sum_{k=1}^K [a_k - E(a_k | \beta)] \quad (6.142)$$

Table 6.3 Newton-Raphson iterative solution for the conditional *MLE* for the ulcer clinical trial data of Example 4.1

i	$\widehat{OR}^{(i)} = e^{\widehat{\beta}^{(i)}}$	$U(\widehat{\beta}^{(i)})$	$U'(\widehat{\beta}^{(i)})$
0	1.0	6.09378	-12.3594
1	1.49305	1.14934	-12.2682
2	1.58673	0.40382	-12.2314
\vdots	\vdots	\vdots	\vdots
11	1.64001	0.00021	-12.2091
12	1.64003	0.00008	-12.2090

and the information is

$$\begin{aligned}
 i(\beta) &= I(\beta) = \sum_{k=1}^K \frac{\left[\sum_i C_{ik} e^{i\beta}\right] \left[\sum_i i^2 C_{ik} e^{i\beta}\right] - \left[\sum_i i C_{ik} e^{i\beta}\right]^2}{\left[\sum_i C_{ik} e^{i\beta}\right]^2} \quad (6.143) \\
 &= \sum_{k=1}^K \left[\frac{\sum_i i^2 C_{ik} e^{i\beta}}{\sum_i C_{ik} e^{i\beta}} - \frac{\left[\sum_i i C_{ik} e^{i\beta}\right]^2}{\left[\sum_i C_{ik} e^{i\beta}\right]^2} \right] \\
 &= \sum_{k=1}^K E[a_k^2|\beta] - E(a_k|\beta)^2 = \sum_{k=1}^K V(a_k|\beta),
 \end{aligned}$$

where $C_{ik} = \binom{n_{1k}}{i} \binom{n_{2k}}{m_{1k} - i}$.

Consider the stratified analysis of the ulcer clinical trial presented in Example 4.1. Applying (6.125) recursively using these expressions provides the *MLE* and its estimated variance. The computations are shown in Table 6.3. Using the null hypothesis value as the starting point ($\beta = 0$), the solution required 12 iterations to reach a solution to five places. The resulting estimate of the *MLE* is $\widehat{\beta} = 0.49472$ with estimated variance $\widehat{V}(\widehat{\beta}) = 0.0819$. The estimated common odds ratio is $\widehat{OR} = 1.64$. These are close to the *MVLE* estimates presented in Table 4.3 for which $\log(\widehat{OR}) = 0.493$ with estimated variance 0.0854 and $\widehat{OR} = 1.637$.

6.8 PROBLEMS

6.1 Consider a sample of N observations where the number of events is distributed as binomial with $P(X) = B(x; N, \pi)$ as in (6.1).

6.1.1. Show that the score equation for the binomial probability is $U(\pi)$ presented in (6.3).

- 6.1.2.** Show that the *MLE* of π is $p = x/N$,
- 6.1.3.** Show the expected information is $I(\pi)$ presented in (6.6).
- 6.1.4.** Show that $i(\hat{\pi}) = I(\hat{\pi})$ in (6.8).
- 6.1.5.** Show that the estimated large sample variance is estimated as $\hat{V}(\hat{\pi}) = p(1 - p)/N$.

6.2 Consider a 2×2 table with a product-binomial likelihood and assume that a logit model applies as in (6.16).

- 6.2.1.** Show that the likelihood and log likelihood are as presented in (6.19) and (6.20), respectively.
- 6.2.2.** Then show that $U(\theta)_\alpha = m_1 - n_1\pi_1 - n_2\pi_2$ in terms of α and β .
- 6.2.3.** Likewise, show that $U(\theta)_\beta = a - n_1\pi_1$.
- 6.2.4.** In addition to the above derivations, show that the score equations can be obtained using the chain rule such as

$$\begin{aligned} U(\theta)_\alpha &= \frac{\partial \ell}{\partial \pi_1} \frac{\partial \pi_1}{\partial \alpha} + \frac{\partial \ell}{\partial \pi_2} \frac{\partial \pi_2}{\partial \alpha} \\ U(\theta)_\beta &= \frac{\partial \ell}{\partial \pi_1} \frac{\partial \pi_1}{\partial \beta}. \end{aligned} \quad (6.144)$$

6.2.5. Likewise, show that the $H(\theta)_\beta$ element of the Hessian matrix can be obtained as

$$H(\theta)_\beta = \frac{\partial}{\partial \beta} U(\theta)_\beta = \frac{\partial}{\partial \beta} \left(\frac{\partial \ell}{\partial \pi_1} \frac{\partial \pi_1}{\partial \beta} \right), \quad (6.145)$$

which is of the common form $d(uv)/d\beta$. Therefore,

$$H(\theta)_\beta = \left[\frac{\partial \ell}{\partial \pi_1} \frac{\partial^2 \pi_1}{\partial \beta^2} + \frac{\partial \pi_1}{\partial \beta} \frac{\partial}{\partial \beta} \left(\frac{\partial \ell}{\partial \pi_1} \right) \right] = \left[\frac{\partial \ell}{\partial \pi_1} \frac{\partial^2 \pi_1}{\partial \beta^2} + \frac{\partial \pi_1}{\partial \beta} \frac{\partial^2 \ell}{\partial \pi_1^2} \frac{\partial \pi_1}{\partial \beta} \right]. \quad (6.146)$$

Solve for the other elements in like fashion.

- 6.2.6.** Jointly solving the two score equations in two unknowns, show that the *MLEs* are $\hat{\alpha} = \log(b/d)$ and $\hat{\beta} = \log(ad/bc)$.

- 6.2.7.** Show that these estimates yield the value zero for the score equations presented in Problems 6.2.2 and 6.2.3.

- 6.2.8.** Then show that the expected information matrix is $\mathbf{I}(\theta) = \mathbf{I}(\alpha, \beta)$, as shown in (6.30).

- 6.2.9.** Show that the large sample covariance matrix of the estimates has elements as shown in (6.36).

- 6.2.10.** From this, show that the estimate of the variance of $\hat{\beta}$ without restrictions, or $\hat{V}(\hat{\beta}|H_1)$, is as shown in (6.37) and that this equals Woolf's estimate of the variance of the log odds ratio under the alternative hypothesis in (2.47).

- 6.2.11.** Then show that the Wald test for the null hypothesis $H_0: \beta = 0$ is as shown in (6.39).

- 6.2.12.** For the single 2×2 table used in Problem 2.11, compute the Wald test and compare it to the Mantel and Cochran tests computed therein.

6.2.13. Evaluating the log likelihood under the null and alternative hypotheses as shown in Section 6.2.3.2, derive the expression for the likelihood ratio test in the 2×2 table presented in (2.97).

6.2.14. Evaluating the score vector under the null hypothesis, show that the score equation for β is as presented in (6.45) and that the estimated information matrix is as presented in (6.47).

6.2.15. Then show that the inverse information matrix is as presented in (6.48) and that the efficient score test is equivalent to Cochran's test.

6.3 Consider a 2×2 table with fixed margins for which the likelihood function is the conditional hypergeometric probability in (2.60).

6.3.1. Derive the score equation $U(\beta)$ in (6.53) and the expression for the information in (6.56).

6.3.2. Then show that (6.57) applies.

6.3.3. Under the null hypothesis $H_0: \beta = 0$, derive the score equation $U(\beta_0)$ in (6.58) and the expression for the information in (6.59).

6.3.4. Then show that the efficient score test equals the Mantel-Haenszel test of Section 2.6.2 for a single 2×2 table as presented in (6.60).

6.3.5. Show that the score-based estimate of the log odds ratio is

$$\hat{\beta} = \left(a - \frac{m_1 n_1}{N} \right) \left(\frac{N^2(N-1)}{m_1 m_2 n_1 n_2} \right) \quad (6.147)$$

and that the variance is as expressed in (6.72).

6.3.6. Compute this estimate, its variance, and a 95% C.I. for the OR for the 2×2 table used in Problem 2.11.

6.4 For a 2×2 table with a product binomial likelihood as in Section 6.2, now assume a log risk model where $\log(\pi_1) = \alpha + \beta$ and $\log(\pi_2) = \alpha$ so that $\beta = \log(\pi_1/\pi_2)$ is the log relative risk.

6.4.1. Show that

$$U(\theta)_\alpha = m_1 - \frac{c\pi_1}{1-\pi_1} - \frac{d\pi_2}{1-\pi_2} \quad (6.148)$$

in terms of α and β .

6.4.2. Likewise, show that

$$U(\theta)_\beta = a - \frac{c\pi_1}{1-\pi_1}. \quad (6.149)$$

6.4.3. Solving jointly, show that the MLEs are $\hat{\alpha} = \log(b/n_2)$ and $\hat{\beta} = \log(an_2/bn_1) = \log(\widehat{RR})$.

6.4.4. Then show that the expected information matrix is

$$\mathbf{I}(\theta) = \mathbf{I}(\alpha, \beta) = \begin{bmatrix} n_1\psi_1 + n_2\psi_2 & n_1\psi_1 \\ n_1\psi_1 & n_1\psi_1 \end{bmatrix}, \quad (6.150)$$

where $\psi_i = \pi_i/(1 - \pi_i)$, $i = 1, 2$.

6.4.5. Also show that the large sample covariance matrix of the estimates has elements

$$\mathbf{V}(\hat{\alpha}, \hat{\beta}) = \mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \frac{1 - \pi_2}{n_2\pi_2} & -\frac{1 - \pi_2}{n_2\pi_2} \\ -\frac{1 - \pi_2}{n_2\pi_2} & \frac{1 - \pi_1}{n_1\pi_1} + \frac{1 - \pi_2}{n_2\pi_2} \end{bmatrix}. \quad (6.151)$$

6.4.6. Then show that the estimated variance of the log relative risk equals that presented in (2.38).

6.4.7. Evaluating the score vector under the null hypothesis $H_0: \beta = \beta_0 = 0$, show that the estimate of α under H_0 implies that $\hat{\pi} = m_1/N$, where $\hat{\pi}$ is the estimated common parameter in both groups.

6.4.8. Then show that the score statistic for β is

$$U(\hat{\boldsymbol{\theta}}_0)_\beta = a - \frac{cm_1}{m_2}. \quad (6.152)$$

Note that H_0 implies that $a:c = m_1:m_2$ and that $E[U(\hat{\boldsymbol{\theta}}_0)_\beta] = 0$.

6.4.9. Then evaluate the estimated information matrix under H_0 and show that the β -element of the inverse estimated information matrix is

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^\beta = \frac{Nm_2}{n_1n_2m_1}. \quad (6.153)$$

6.4.10. Show that the efficient score test is

$$X^2 = \left(a - \frac{cm_1}{m_2} \right)^2 \left(\frac{Nm_2}{n_1n_2m_1} \right). \quad (6.154)$$

6.4.11. For the single 2×2 table used in Problem 2.11, compute the score test and compare it to the tests computed previously.

6.4.12. Show that the score-based estimate of the common log relative risk is

$$\hat{\beta} = \log(\widehat{RR}) = \left(a - \frac{cm_1}{m_2} \right) \left(\frac{Nm_2}{n_1n_2m_1} \right) \quad (6.155)$$

with estimated variance

$$\hat{V}(\hat{\beta}) = \frac{Nm_2}{n_1n_2m_1}. \quad (6.156)$$

6.4.13. For the data in Problem 2.11, compute the score-based estimate of the $\log(RR)$ and its large sample variance and compare these to the simple estimates computed in Problem 2.11.

6.5 Now consider a set of K independent 2×2 tables. Conditioning on the margins of each table as in Section 6.5.1, the model assumes a conditional hypergeometric likelihood for each table with a common odds ratio $OR_k = \varphi = e^\beta$.

6.5.1. From the likelihood in (6.52), derive the score equation $U(\beta)$ in (6.142) and the information $I(\beta)$ in (6.143).

6.5.2. Under $H_0: \beta = 0$, show that the score equation for β is as shown in (6.74) and the information is as shown in (6.75).

6.5.3. Show that the efficient score test of H_0 in (6.76) equals the Mantel-Haenszel test for stratified 2×2 tables presented in Section 4.2.3.

6.5.4. Show that the score-based stratified-adjusted estimate of the common log odds ratio in (6.77) is

$$\hat{\beta} = \sum_k \left(a_k - \frac{m_{1k}n_{1k}}{N_k} \right) \left(\sum_l \frac{m_{1l}m_{2l}n_{1l}n_{2l}}{N_l^2(N_l - 1)} \right)^{-1} \quad (6.157)$$

and that

$$V(\hat{\beta}) = \left(\sum_k \frac{m_{1k}m_{2k}n_{1k}n_{2k}}{N_k^2(N_k - 1)} \right)^{-1}. \quad (6.158)$$

6.5.5. For the stratified 2×2 tables used in Problem 4.9, compute this estimate, its variance, and a 95% C.I. for the common OR .

6.6 Consider the case of K independent 2×2 tables with a compound product binomial likelihood. Section 6.2.2 presents the score equations and Hessian for a single 2×2 table using a logit model with parameters α and β . Use these with the compound product binomial likelihood for K such tables presented in (6.79) and (6.80) in terms of the parameters $\theta = (\alpha_1 \cdots \alpha_K \beta)^T$.

6.6.1. Show that

$$U(\theta)_{\alpha_k} = \frac{\partial \ell_k}{\partial \alpha_k} = m_{1k} - n_{1k} \frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} - n_{2k} \frac{e^{\alpha_k}}{1 + e^{\alpha_k}} \quad (6.159)$$

for $k = 1, \dots, K$; and that

$$U(\theta)_{\beta} = \frac{\partial \ell}{\partial \beta} = \sum_{k=1}^K \left(a_k - n_{1k} \frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} \right). \quad (6.160)$$

6.6.2. Show that the elements of the Hessian matrix then are

$$\mathbf{H}(\theta)_{\alpha_k} = \frac{\partial^2 \ell_k}{\partial \alpha_k^2} = -n_{1k} \left[\frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} - \left(\frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} \right)^2 \right] \quad (6.161)$$

$$- n_{2k} \left[\frac{e^{\alpha_k}}{1 + e^{\alpha_k}} - \left(\frac{e^{\alpha_k}}{1 + e^{\alpha_k}} \right)^2 \right]$$

$$\mathbf{H}(\theta)_{\beta} = \frac{\partial^2 \ell}{\partial \beta^2} = - \sum_{k=1}^K n_{1k} \left[\frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} - \left(\frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} \right)^2 \right]$$

$$\mathbf{H}(\theta)_{\alpha_k \beta} = \frac{\partial^2 \ell}{\partial \alpha_k \partial \beta} = -n_{1k} \left[\frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} - \left(\frac{e^{\alpha_k + \beta}}{1 + e^{\alpha_k + \beta}} \right)^2 \right].$$

6.6.3. Under the hypothesis of no association in all of the tables, $H_0: \beta = \beta_0 = 0$, show that $\hat{\alpha}_{k(0)} = \log(m_{1k}/m_{2k})$ and that the score equation for β is as presented in (6.86).

6.6.4. Then derive the elements of the estimated information matrix as presented in (6.87).

6.6.5. Also derive the expression for $\mathbf{I}(\hat{\theta}_0)^\beta$ under H_0 in (6.89).

6.6.6. Show that the efficient score statistic is Cochran's test as shown in (6.90).

6.7 Similarly, assume that there is a common relative risk among the K 2×2 tables represented by a log link as in Problem 6.4, where $\beta = \log(RR)$.

6.7.1. Under $H_0: \beta = \beta_0 = 0$, show that the score equation for α_k is

$$U(\theta_0)_{\alpha_k} = m_{1k} - m_{2k} \frac{e^{\alpha_k}}{1 + e^{\alpha_k}}$$

and that the solution is

$$\hat{\alpha}_{k(0)} = \log(m_{1k}/N_k).$$

6.7.2. Under H_0 , show that the score statistic for β is

$$U(\hat{\theta}_0)_\beta = \sum_k a_k - \frac{c_k m_{1k}}{m_{2k}}. \quad (6.162)$$

6.7.3. Then evaluate the inverse information matrix under H_0 and show that the β -element of the inverse information matrix is

$$\mathbf{I}(\hat{\theta}_0)^\beta = \left(\sum_k \frac{n_{1k} n_{2k} m_{1k}}{N_k m_{2k}} \right)^{-1}, \quad (6.163)$$

where $\hat{\pi}_k = m_{1k}/N_k$ is the estimated common probability in the k th stratum.

6.7.4. Show that the resulting score test is

$$X^2 = \frac{\left[\sum_k a_k - \frac{c_k m_{1k}}{m_{2k}} \right]^2}{\sum_k \frac{n_{1k} n_{2k} m_{1k}}{N_k m_{2k}}}. \quad (6.164)$$

6.7.5. For comparison, show that the above expression is equivalent to that for the Radhakrishna asymptotically fully efficient test for a common relative risk presented in (4.188) of Problem 4.4.3 based on (4.97) of Section 4.7.2. Gart (1985) presented an alternative derivation of the score test and its equivalence to the Radhakrishna test using a slightly different parameterization.

6.8 Consider Cox's (1958b) logistic model for matched pairs described in Section 6.6.1.

6.8.1. Starting with the unconditional likelihood described in Section 6.6.1, derive the expressions for the likelihood in (6.95) and log likelihood in (6.96).

6.8.2. Then derive the score equations for the pair-specific effects $U(\theta)_{\alpha_i}$ in (6.97) and the common log odds ratio $U(\theta)_\beta$ in (6.98).

6.8.3. Then, conditioning on $S_i = (y_{i1} + y_{i2}) = 1$, show that the conditional likelihood in (6.103) can be expressed as

$$L(\beta)_{|S=1} = \prod_{i:S_i=1} \frac{[\pi_{i1}(1-\pi_{i2})]^{y_{i1}(1-y_{i2})} [\pi_{i2}(1-\pi_{i1})]^{y_{i2}(1-y_{i1})}}{\pi_{i1}(1-\pi_{i2}) + \pi_{i2}(1-\pi_{i1})}. \quad (6.165)$$

6.8.4. Then show that the expressions in (6.104) and (6.105) hold.

6.8.5. Show that this yields $L(\beta)_{|S=1}$ as in (6.106).

6.8.6. Then derive $\ell_{(c)}$ in terms of f and $M = f + g$.

6.8.7. Derive $U(\beta)$ in (6.107) and show that $\hat{\beta} = \log(f/g)$.

6.8.8. Then show that $i(\beta)$ is as expressed in (6.109) and $I(\beta)$ as in (6.110).

6.8.9. Then show that $V(\hat{\beta})$ is as expressed in (6.112).

6.8.10. Under $H_0: \beta = 0$, derive the likelihood ratio test presented in (6.115).

6.8.11. Derive the efficient score test statistic in Section 6.6.4 and show that it equals McNemar's statistic.

6.9 Consider the recombination fraction example presented in Section 6.7.1.

6.9.1. Based on the assumed probabilities, show that a simple moment estimator of the recombination fraction is $\hat{\theta}^{(0)} = (x_1 - x_2 - x_3 + x_4)/N$.

6.9.2. Either write a computer program or use a calculator to fit the model using Newton-Raphson iteration convergent to eight decimal places. Verify the computations presented in Table 6.1

6.9.3. Do likewise, using Fisher scoring to verify the computations in Table 6.2.

6.9.4. Using the final $MLE \hat{\theta}$ and estimated observed information $i(\hat{\theta})$ from the Newton-Raphson iteration, which provides an estimate of the variance of the estimate, compute the 95% *C.I.* for θ .

6.9.5. Likewise, using the final estimated expected information from Fisher scoring, compute a 95% *C.I.* for θ .

6.9.6. For $H_0: \theta = \theta_0 = 0.25$ compute the Wald, score, and likelihood ratio tests.

6.10 Consider the estimation of the conditional *MLE* of the log odds ratio β based on the conditional hypergeometric distribution.

6.10.1. For a single 2×2 table, use the score equation in (6.53) and the observed information in (6.56) as the basis for a computer program to use the Newton-Raphson iteration to solve for the *MLE* of the log odds ratio β and to then evaluate the estimated expected information $I(\hat{\beta})$.

6.10.2. Apply this program to the data from Problem 2.11 and compare the point estimate of the log odds ratio and its estimated variance, and the 95% asymmetric confidence limits for the odds ratio, to those computed previously.

6.10.3. Also, generalize this program to allow for K independent strata as in Section 6.5.1 and Example 6.7.

6.10.4. Apply this program to the data from Example 4.1 to verify the computations in Table 6.3.

6.11 Now consider the maximum likelihood estimate of the common odds ratio in K independent 2×2 tables based on the compound product-binomial likelihood of Section 6.5.2 and Problem 6.6.

6.11.1. Use the expressions for the score equations and the Hessian matrix to write a program for Newton-Raphson iteration to jointly solve for the *MLE* of the vector θ .

6.11.2. Apply the Newton-Raphson iteration to the set of 2×2 tables in Problem 4.9 to obtain the *MLE* of the common log odds ratio and its estimated variance, and the 95% asymmetric confidence limits for the odds ratio, and compare these to those computed previously.

Logistic Regression Models

Chapter 6 describes the analysis of 2×2 tables using a simple logit model. This model provides estimates and tests of the log odds ratio for the probability of the positive response or outcome of interest as a function of the exposure or treatment group. We now consider the simultaneous estimation of the effects of multiple covariates, perhaps some categorical and some quantitative, on the odds of the response using a logistic regression model. We first consider the logistic regression model for independent observations, as in a cross-sectional, prospective, or retrospective study. The treatment herein is based on the work of Cox (1970) and others, using maximum likelihood as the basis for estimation. Later we consider the conditional logistic model for paired or clustered observations, as in matched sets, described by Breslow (1982). We then consider generalizations to a polychotomous dependent variable using a multinomial logistic model for nominal scale outcomes, and a proportional odds model for ordinal outcomes. We also provide an introduction to mixed models that include random effects, and to multivariate or longitudinal models based on generalized estimating equations (GEE). An excellent general references are the books by Hosmer and Lemeshow (2004) and Stokes et al. (2000).

7.1 UNCONDITIONAL LOGISTIC REGRESSION MODEL

7.1.1 General Logistic Regression Model

In keeping with the notation in a single 2×2 table, we start with a prospective cohort or cross-sectional sample of N observations, of which m_1 have the positive response (e.g., develop the disease) and m_2 do not. For each subject, the response is represented by the binary dependent variable Y , where for the i th subject, $y_i = 1$

if a positive response is observed, 0 if not. We assume that the probability of the positive response can be expressed as a linear function of a p covariate vector $\mathbf{X} = (X_1 \cdots X_p)^T$ and the $p + 1$ parameter vector $\boldsymbol{\theta} = (\alpha \ \beta_1 \ \cdots \ \beta_p)^T$. Then for the i th subject, $\pi_i(\mathbf{x}_i, \boldsymbol{\theta}) = P(y_i = 1 | \mathbf{x}_i, \boldsymbol{\theta})$ as a function of that subject's covariate vector $\mathbf{x}_i = (x_{i1} \cdots x_{ip})^T$. The covariates can be either quantitative characteristics or design effects to represent qualitative characteristics. We then assume that the logit of these probabilities can be expressed as a function of the p covariates, such as a linear function of the parameters $\boldsymbol{\theta}$ of the form

$$\log \left[\frac{\pi_i}{1 - \pi_i} \right] = \alpha + \mathbf{x}'_i \boldsymbol{\beta} = \alpha + \sum_{j=1}^p x_{ij} \beta_j, \quad (7.1)$$

where $\pi_i = \pi_i(\mathbf{x}_i, \boldsymbol{\theta})$ and $\boldsymbol{\beta} = (\beta_1 \ \cdots \ \beta_p)^T$. This implies that the probabilities can be obtained as the logistic function (inverse logit)

$$\begin{aligned} \pi_i &= \frac{e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}}{1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}} = \frac{1}{1 + e^{-(\alpha + \mathbf{x}'_i \boldsymbol{\beta})}} \\ 1 - \pi_i &= \frac{1}{1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}}. \end{aligned} \quad (7.2)$$

Since the i th observation is a Bernoulli variable $\{y_i\}$, then the likelihood in terms of the probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)$ is

$$L(\boldsymbol{\pi}) = \prod_{i=1}^N \pi_i^{y_i} (1 - \pi_i)^{1-y_i}. \quad (7.3)$$

Expressing the probabilities as a logistic function of the covariates and the parameters yields

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \left(\frac{e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}}{1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}} \right)^{1-y_i}. \quad (7.4)$$

Likewise, the log likelihood in terms of the probabilities is

$$\ell(\boldsymbol{\pi}) = \sum_{i=1}^N y_i \log \pi_i + \sum_{i=1}^N (1 - y_i) \log (1 - \pi_i), \quad (7.5)$$

and that in terms of the corresponding model parameters $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N y_i (\alpha + \mathbf{x}'_i \boldsymbol{\beta}) - \sum_{i=1}^N \log \left(1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} \right). \quad (7.6)$$

The score vector for the parameters is

$$\mathbf{U}(\boldsymbol{\theta}) = \left[U(\boldsymbol{\theta})_\alpha \ U(\boldsymbol{\theta})_{\beta_1} \ \cdots \ U(\boldsymbol{\theta})_{\beta_p} \right]^T. \quad (7.7)$$

The score equation for the intercept is

$$\begin{aligned} U(\boldsymbol{\theta})_\alpha &= \frac{\partial \ell}{\partial \alpha} = \sum_i y_i - \sum_i \frac{e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}}{1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}} \\ &= \sum_i (y_i - \pi_i) = \sum_i [y_i - E(y_i | \mathbf{x}_i)] = m_1 - E(m_1 | \boldsymbol{\theta}), \end{aligned} \quad (7.8)$$

where $E(y_i | \mathbf{x}_i) = \pi_i$. The score equation for the j th coefficient is

$$\begin{aligned} U(\boldsymbol{\theta})_{\beta_j} &= \frac{\partial \ell}{\partial \beta_j} = \sum_i x_{ij} \left(y_i - \frac{e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}}{1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}} \right) \\ &= \sum_i x_{ij} [y_i - \pi_i] = \sum_i x_{ij} [y_i - E(y_i | \mathbf{x}_i)], \end{aligned} \quad (7.9)$$

which is a weighted sum of the observed value of the response (y_i) minus that expected under the model.

Note that each score equation involves the complete parameter vector $\boldsymbol{\theta}$. The *MLE* is the vector $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha} \ \widehat{\beta}_1 \ \dots \ \widehat{\beta}_p)^T$ that jointly satisfies $U(\boldsymbol{\theta}) = \mathbf{0}$; or the solution obtained by setting the $p + 1$ score equations to zero and solving simultaneously. Since closed-form solutions do not exist, the solution must be obtained by an iterative procedure such as the Newton-Raphson algorithm.

Note that an important consequence of the score equation for the intercept is that setting $U(\boldsymbol{\theta})_\alpha = 0$ requires that the *MLE* for α and the estimates of the coefficients jointly satisfy $U(\widehat{\boldsymbol{\theta}})_\alpha = m_1 - \sum_i \widehat{\pi}_i = 0$, where $\widehat{\pi}_i$ is the estimated probability obtained upon substituting $\widehat{\alpha} + \mathbf{x}'_i \widehat{\boldsymbol{\beta}}$ into (7.2). Thus, the mean estimated probability is $\widehat{\pi} = \sum_i \widehat{\pi}_i / N = m_1 / N$.

The observed information $\mathbf{i}(\boldsymbol{\theta}) = -\mathbf{H}(\boldsymbol{\theta})$ has the following elements. For $\mathbf{i}(\boldsymbol{\theta})_\alpha$ we require

$$\frac{\partial \pi_i}{\partial \alpha} = \frac{e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}}{(1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}})^2} = \pi_i (1 - \pi_i), \quad (7.10)$$

so that

$$\mathbf{i}(\boldsymbol{\theta})_\alpha = \frac{-\partial U(\boldsymbol{\theta})_\alpha}{\partial \alpha} = \sum_i \pi_i (1 - \pi_i). \quad (7.11)$$

Likewise, it is readily shown that the remaining elements of the observed information matrix for $1 \leq j < k \leq p$ are

$$\begin{aligned} \mathbf{i}(\boldsymbol{\theta})_{\beta_j} &= \frac{-\partial U(\boldsymbol{\theta})_{\beta_j}}{\partial \beta_j} = \sum_i x_{ij} \frac{\partial \pi_i}{\partial \beta_j} = \sum_i x_{ij} \frac{x_{ij} e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}}{(1 + e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}})^2} \\ &= \sum_i x_{ij}^2 \pi_i (1 - \pi_i), \end{aligned} \quad (7.12)$$

$$\mathbf{i}(\boldsymbol{\theta})_{\beta_j, \beta_k} = \frac{-\partial U(\boldsymbol{\theta})_{\beta_j}}{\partial \beta_k} = \sum_i x_{ij} \frac{\partial \pi_i}{\partial \beta_k} = \sum_i x_{ij} x_{ik} \pi_i (1 - \pi_i), \quad (7.13)$$

$$\mathbf{i}(\boldsymbol{\theta})_{\alpha, \beta_j} = \frac{-\partial U(\boldsymbol{\theta})_\alpha}{\partial \beta_j} = \sum_i \frac{\partial \pi_i}{\partial \beta_j} = \sum_i x_{ij} \pi_i (1 - \pi_i). \quad (7.14)$$

Since $\mathbf{i}(\boldsymbol{\theta})$ can be expressed in terms of the probabilities $\{\pi_i\}$ or the parameters $\{\boldsymbol{\theta}\}$, then $\mathbf{i}(\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})$ and the observed and expected information are the same for this model. Also, since $V(y_i | \mathbf{x}_i) = \pi_i (1 - \pi_i)$ under the model, then the elements of the information matrix are weighted sums of the model-based Bernoulli variances, such as $\mathbf{I}(\boldsymbol{\theta})_{\beta_j} = \sum_i x_{ij}^2 V(y_i | \mathbf{x}_i)$, where $V(y_i | \mathbf{x}_i) = \pi_i (1 - \pi_i)$.

Using the elements of the Hessian matrix, the vector of parameter estimates can be solved using the Newton-Raphson iterative procedure, or another algorithm. Wald tests and large sample confidence limits for the individual parameters, such as the j th coefficient β_j , are readily computed using the large sample variance $\widehat{V}(\widehat{\beta}_j) = [\mathbf{I}(\widehat{\theta})^{-1}]_{\beta_j}$ obtained as the corresponding diagonal element of the estimated expected information, $\mathbf{I}(\widehat{\theta})$. The null and fitted likelihoods can also be used as the basis for a likelihood ratio test of the model, as described subsequently.

Section 6.2.2 shows that the estimated coefficient in a simple logit model for a 2×2 table is the *MLE* of the log odds ratio. Subsequently (see Section 7.2.1), we show that the estimated coefficients for binary covariates have the same interpretation in a multivariate logistic regression model. That is, if X_j is coded as 1 or 0, then the estimated odds ratio for this covariate is $\widehat{OR}_j = \exp(\widehat{\beta}_j)$. Confidence limits on the coefficient then yield asymmetric confidence limits on the odds ratio.

In some applications it is also of interest to obtain confidence limits for the predicted probability for an observation with covariate vector \mathbf{x} . Let $\tilde{\mathbf{x}}$ denote the covariate vector augmented by the constant for the intercept $\tilde{\mathbf{x}} = (1 // \mathbf{x})$. Then the linear predictor is

$$\widehat{\eta} = \tilde{\mathbf{x}}' \widehat{\theta}. \quad (7.15)$$

Denoting the estimated variance of the estimates as $\widehat{\Sigma}_{\theta} = \widehat{V}(\widehat{\theta}) = \mathbf{I}(\widehat{\theta})^{-1}$, the estimated variance of the linear predictor is

$$\widehat{V}(\widehat{\eta}) = \tilde{\mathbf{x}}' \widehat{\Sigma}_{\theta} \tilde{\mathbf{x}} = \widehat{\sigma}_{\widehat{\eta}}^2. \quad (7.16)$$

Therefore, the $(1 - \alpha)$ -level confidence limits on η are

$$(\widehat{\eta}_l, \widehat{\eta}_u) = \widehat{\eta} \pm Z_{1-\alpha/2} \widehat{\sigma}_{\widehat{\eta}}, \quad (7.17)$$

and the resulting confidence limits on the true probability π are

$$\frac{e^{\widehat{\eta}_l}}{1 + e^{\widehat{\eta}_l}} \leq \pi \leq \frac{e^{\widehat{\eta}_u}}{1 + e^{\widehat{\eta}_u}}. \quad (7.18)$$

7.1.2 Logistic Regression and Binomial Logit Regression

Section 6.2.2 describes the logit model for the probabilities of a positive response or of developing the disease (D) in a single 2×2 table. That model is a special case of the more general model above. Consider the case where X is a single binary covariate and for the i th subject $x_i = 1$ if E , or 0 if \bar{E} , for exposed versus not exposed. Then the data can be summarized in a 2×2 table with frequencies.

		x_i		N
		(1)	(0)	
y_i	E	a	b	m_1
	\bar{E}	c	d	m_2
		n_1	n_2	

Then the logistic regression likelihood is

$$\begin{aligned}
 L &= \prod_{i=1}^N \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \prod_{i=1}^N \left(\frac{e^{\alpha+x_i\beta}}{1+e^{\alpha+x_i\beta}} \right)^{y_i} \left(\frac{1}{1+e^{\alpha+x_i\beta}} \right)^{1-y_i} \\
 &= \left(\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} \right)^a \left(\frac{1}{1+e^{\alpha+\beta}} \right)^c \left(\frac{e^\alpha}{1+e^\alpha} \right)^b \left(\frac{1}{1+e^\alpha} \right)^d \\
 &= \pi_1^a (1 - \pi_1)^c \pi_2^b (1 - \pi_2)^d,
 \end{aligned} \tag{7.19}$$

which equals the product binomial logit model likelihood in (6.19), where $\pi_1 = P(D|E) = \pi_{i:x_i=1}$ and $\pi_2 = P(D|\bar{E}) = \pi_{i:x_i=0}$.

Thus, for a simple qualitative covariate, such as E versus \bar{E} , an equivalent model can be obtained using the individual covariate values for all subjects, or using a product binomial logit model. The latter approach is also called binomial logit regression, or simply binomial regression.

Binomial regression also generalizes to a higher-order contingency table layout where the covariate vector for each subject, x_i , is now a vector of binary covariates, as illustrated in the following example.

Example 7.1 Coronary Heart Disease in the Framingham Study

Cornfield (1962) presents preliminary results from the Framingham study of the association of cholesterol levels and blood pressure with the risk of developing coronary heart disease (CHD) in the population of the town of Framingham, Massachusetts over a six-year period. These data were described previously in Examples 2.20 and 2.21. Cholesterol levels and blood pressure were measured at the beginning of the study and the cohort was then followed over time. For the purpose of illustration, these predictor variables have been categorized as high versus low values of serum cholesterol (X_1 : $hichol = 1$ if ≥ 200 mg/dL, 0 if < 200 mg/dL) and as high versus low values of systolic blood pressure (X_2 : $hisbp = 1$ if ≥ 147 mmHg, 0 if < 147 mmHg). Thus, there are four possible configurations of the covariates, and for each configuration ($j = 1, \dots, 4$) the data consist of (a_j, n_j, \mathbf{x}_j) , where a_j is the number of patients developing CHD among the n_j subjects enrolled into the cohort (or the number at risk) for subjects sharing the covariate vector $\mathbf{x}_j = (x_{1j} \ x_{2j})^T$. From the sample of $N = 1329$ subjects, the data in each category of the covariates are presented in Table 7.1.

A logistic or binomial regression model could then be fit using a statistical package such as PROC LOGISTIC in SAS with a data set containing two observations per stratum. For example, for the first stratum ($j = 1$) with two cells ($l = 1, 2$), the data are represented as

Cell (l)	x_l	y_l	F_l
1	1 0 0	1	10
2	1 0 0	0	421

(see Example 7.2). For each stratum or configuration of the covariates, the first observation specifies the number (F) with the response ($Y = 1$) and the second

Table 7.1 Incidence of coronary heart disease in the Framingham study as a function of blood pressure and cholesterol levels.

Category <i>j</i>	# CHD <i>a_j</i>	# Subjects <i>n_j</i>	(Intercept, Covariates) (1, x_{1j} , x_{2j})
1	10	431	(1, 0, 0)
2	10	142	(1, 0, 1)
3	38	532	(1, 1, 0)
4	34	224	(1, 1, 1)

specifies the number (F) without the response ($Y = 0$). In this case the frequency F is also the weight for each term in the likelihood, where

$$L = \prod_l \pi_l^{F_l y_l} (1 - \pi_l)^{F_l (1 - y_l)}, \quad (7.20)$$

$l = 1, \dots, 8$ in this case, and

$$\ell = \sum_l F_l y_l [\log \pi_l] + \sum_l F_l (1 - y_l) \log (1 - \pi_l), \quad (7.21)$$

where $F_l y_l = a_j$ and $F_l (1 - y_l) = (n_j - a_j)$ for the j th corresponding stratum or category in Table 7.1. This is the same as the likelihood for 1329 individual observations with one record per subject.

An iterative solution is then required to fit the model. It is customary to start with an initial value of the coefficient vector evaluated under the null hypothesis, that is, using $\hat{\theta}^{(0)} = 0$ for the coefficients. The initial value for the intercept may be set to that for the simple logit model for a 2×2 table in (6.43): $\hat{\alpha}^{(0)} = \log [m_1/m_2]$. For this example $m_1/(N - m_1) = 92/1237$ and $\hat{\alpha}^{(0)} = -2.5987$. The solution converges to eight places after only six iterations, as shown in Table 7.2. Since $I(\theta) = i(\theta)$ for logistic regression, then Newton-Raphson and Fisher scoring are equivalent. Thus, $I(\hat{\theta})^{-1}$ at the final iteration provides the estimated covariance matrix of the coefficients.

Note that this is a main effects only model. Since there are four covariate categories, the saturated model would have three coefficients in addition to the intercept, the missing term being the interaction between the two covariates. This interaction model is described in Section 7.4.

7.1.3 SAS Procedures

Most statistical applications provide a program for the computation of logistic regression models. In SAS, PROC CATMOD was one of the earliest such procedures. It implements the method of Grizzle et al. (1969), which fits generalizations of the binomial logit model. PROC LOGISTIC is a more general, comprehensive procedure for fitting logistic regression models. PROC GENMOD, fits the wide family

Table 7.2 Newton-Raphson iterative solution of logit model for data in Table 7.1.

i		$\hat{\theta}$	$U(\hat{\theta})$	$i(\hat{\theta})^{-1} = -H(\hat{\theta})^{-1}$		
0	$\hat{\alpha}^{(0)}$	-2.5986560	0.0000000	0.030690	-0.026380	-0.0145450
	$\hat{\beta}_1^{(0)}$	0.0000000	19.665914	-0.026380	0.047753	-0.0028450
	$\hat{\beta}_2^{(0)}$	0.0000000	18.663657	-0.014545	-0.002845	0.0586903
1	$\hat{\alpha}^{(1)}$	-3.3889060	-16.78430	0.0487293	-0.0396770	-0.0189040
	$\hat{\beta}_1^{(1)}$	0.8859976	-10.34558	-0.0396770	0.0541921	-0.0010600
	$\hat{\beta}_2^{(1)}$	1.0394225	-10.46654	-0.0189040	-0.0010600	0.0416894
6	$\hat{\alpha}$	-3.6282170	-0.000000	0.0612917	-0.0515490	-0.0210170
	$\hat{\beta}_1$	1.0301350	-0.000000	-0.0515490	0.0678787	-0.0014840
	$\hat{\beta}_2$	0.9168417	-0.000000	-0.0210170	-0.0014840	0.0485391

of *generalized linear models (GLMs)* described in Section A.10. These models are based on distributions from the exponential family that includes as a special case the logistic regression model based on the binomial distribution. The GENMOD procedure provides some features not available through PROC LOGISTIC. Then PROC GLIMMIX is a further generalization that allows for the inclusion of random effects into the family of generalized models, including logistic models.

In logistic regression we wish to model the probability of either a positive or negative outcome or characteristic (D versus \bar{D}) by expressing the log odds (logit) of one of these probabilities as a function of the covariates. Herein we employ the usual convention that the data analyst is interested in describing covariate effects on the odds of a positive response (D) through an assessment of the covariate effects on the log odds, say $\log(O_x)$, where the odds of the positive response associated with a given value of the covariate vector x is

$$O_x = \frac{P(D|x)}{P(\bar{D}|x)} = \frac{\pi_x}{1 - \pi_x}. \quad (7.22)$$

In describing the odds in this manner, the negative (\bar{D}) category is considered the reference category or the denominator for calculation of the odds ratio.

The dependent variable Y that designates \bar{D} and D can be coded as any unique pair of values: (0,1) or (1,2), and so on. The default convention in PROC LOGISTIC is that the highest-numbered or last category alphanumerically is used as the reference category. Thus, if Y is coded such that $\bar{D} = 0$ and $D = 1$, or as $\bar{D} = 1$ and $D = 2$,

Table 7.3 SAS program for the Framingham CHD data in Table 7.1.

```

data one; input hichol hisbp chd frequncy; cards;
0 0 1 10
0 0 0 421
0 1 1 10
0 1 0 132
1 0 1 38
1 0 0 494
1 1 1 34
1 1 0 190
;
proc logistic descending;
  model chd = hichol hisbp / rl;
  weight frequncy;
run;

```

then the default in PROC LOGISTIC is to use the highest-numbered category D as the reference category in these cases rather than \bar{D} as in (7.22). In such cases, the *descending* option must be used to ensure that the program uses the category corresponding to the desired \bar{D} as the reference category.

Other useful options include *covb*, which computes the estimated covariance matrix of the estimates, and *rl* (or risk limits), which computes the estimated odds ratios associated with each covariate.

PROC LOGISTIC (and CATMOD) also provide for a polychotomous dependent variable as described in later sections.

Example 7.2 Coronary Heart Disease in the Framingham Study (continued)

For the Example 7.1, the SAS program in Table 7.3 would perform the logistic regression analysis. The SAS output is presented in Table 7.4 (other extraneous information deleted). The program first presents the score and likelihood ratio tests of significance of the overall model, which in this case is the null hypothesis that the two regression coefficients are zero. Each is highly significant. These tests are discussed later. This is followed by the estimates of the model parameters, standard errors, and Wald tests for the individual parameters.

In this example, the *rl* option provides the estimated odds ratio associated with the two binary covariates, each adjusted for the effect of the other. The effect of high versus low cholesterol, adjusting for blood pressure, is $\hat{\beta}_1 = 1.0301$ with an estimated odds ratio of $\widehat{OR}_{H:L} = e^{1.0301} = 2.801$. This indicates that high cholesterol is associated with a 180% increase $[(2.801 - 1) \times 100]$ in the odds of CHD (adjusted for the effects of blood pressure). The inverse provides the decrease in odds associated with low versus high cholesterol, or $\widehat{OR}_{L:H} = e^{-1.0301} = 0.357$, which indicates a 64.3% decrease in the odds of CHD $[(1 - 0.357) \times 100]$. Similarly,

$\hat{\beta}_2 = 0.9168$ is the effect of high versus low systolic blood pressure adjusted for serum cholesterol, which yields an odds ratio of $\widehat{OR}_{H:L} = 2.501$, indicating a 150% increase in the odds among those with high SBP. The large sample 95% confidence limits are also presented for the odds ratio associated with each covariate.

7.1.4 Stratified 2×2 Tables

Section 6.5.2 describes the derivation of the Cochran test of association for a common odds ratio in K independent 2×2 tables as a score test in a logit model as described by Day and Byar (1979). In (6.81) the probabilities of a positive response in the j th stratum, $\pi_{1j} = P_j(D|E)$ and $\pi_{2j} = P_j(D|\bar{E})$, are expressed as logistic functions with parameters $(\alpha_1, \dots, \alpha_K, \beta)$. An equivalent model could be fit using PROC LOGISTIC with the covariate vector for the i th subject

$$\mathbf{x}_i = (z_{i2} \ z_{i3} \ \dots \ z_{iK} \ x_i)^T, \quad (7.23)$$

where $z_{ij} = I(\text{subject } i \in \text{stratum } j)$ is a binary indicator variable for membership of the i th subject in the j th stratum, $j = 2, \dots, K$, and $x_i = I(E)$ is the indicator variable for membership in the exposed group, $i = 1, \dots, N$. The associated parameter vector is

$$\boldsymbol{\theta} = (\alpha \ \gamma_2 \ \gamma_3 \ \dots \ \gamma_K \ \beta)^T, \quad (7.24)$$

where α is the stratum 1 effect, stratum 1 being the reference category in this case; γ_j is the difference between the j th stratum and the stratum 1 effects ($j > 1$); and β is the stratified-adjusted effect of exposure that is assumed to be constant for all strata, that is, $\beta = \log(OR)$ as in Chapter 4. In terms of the Day-Byar parameterization in Section 6.5.2, $\alpha_1 \equiv \alpha$ and $\alpha_j \equiv \alpha + \gamma_j$.

The linear predictor is then expressed as

$$\eta = \alpha + \mathbf{x}'_i \boldsymbol{\beta} = \alpha + \sum_{j=2}^K z_{ij} \gamma_j + x_i \beta. \quad (7.25)$$

Therefore,

$$\begin{aligned} \pi_{11} &= \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} & \pi_{21} &= \frac{e^\alpha}{1 + e^\alpha} \\ \pi_{1j} &= \frac{e^{\alpha+\gamma_j+\beta}}{1 + e^{\alpha+\gamma_j+\beta}} & \pi_{2j} &= \frac{e^{\alpha+\gamma_j}}{1 + e^{\alpha+\gamma_j}} \quad 2 \leq j \leq K, \end{aligned} \quad (7.26)$$

Note that γ_j represents the log odds ratio for stratum j versus the reference stratum 1 both for those exposed, and also for those not exposed, that is,

$$\frac{\pi_{1j}/(1 - \pi_{1j})}{\pi_{11}/(1 - \pi_{11})} = \frac{\pi_{2j}/(1 - \pi_{2j})}{\pi_{21}/(1 - \pi_{21})} = e^{\gamma_j}, \quad (7.27)$$

and that the maximum likelihood estimate of the common odds ratio for exposure is $\widehat{OR}_{MLE} = e^{\hat{\beta}}$.

Table 7.4 Logistic regression analysis of the data in Table 7.1.

The LOGISTIC Procedure

Model Information

Response Profile

Ordered Value	chd	Total Frequency	Total Weight
1	1	4	92.0000
2	0	4	1237.0000

Probability modeled is chd=1.

Model Fit Statistics

Criterion	Only	Intercept and Covariates
-2 Log L	668.831	632.287

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	36.5443	2	<.0001
Score	36.8234	2	<.0001
Wald	33.8248	2	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	-3.6282	0.2476	214.7756	<.0001
hichol	1	1.0301	0.2605	15.6334	<.0001
hisbp	1	0.9168	0.2203	17.3180	<.0001

Odds Ratio Estimates

Effect	Estimate	Point	95% Wald	Confidence Limits
hichol	2.801	1.681	4.668	
hisbp	2.501	1.624	3.852	

Example 7.3 *Clinical Trial in Duodenal Ulcers (continued)*

As a problem, the reader is asked to conduct a logistic regression analysis of the data from the ulcer healing clinical trial presented in Example 4.1. The *MLEs* of the odds ratios, the corresponding asymmetric Wald 95% confidence limits, and the Wald test *p*-values for the model effects are

Effect	\widehat{OR}	95% C.I.	$p \leq$
Stratum 2 versus 1	2.305	0.861, 6.172	0.0964
Stratum 3 versus 1	1.388	0.765, 2.519	0.2813
Drug versus Placebo	1.653	0.939, 2.909	0.0816

Neither stratum 2 nor 3 has an effect on the overall risk significantly different from that of the first stratum, which is consistent with the results of the test of stratum by healing presented in Section 4.4.3. The odds of healing are increased 1.65-fold with drug treatment, which is not statistically significant. The estimated common odds ratio for drug treatment versus placebo, adjusted for stratum effects, and its confidence limits, are comparable to the Mantel-Haenszel and *MVLE* estimates presented earlier in Chapter 4, and the conditional *MLE* presented in Example 6.7.

7.1.5 Family of Binomial Regression Models

The logistic regression model is one of a family of binomial regression models for such data that employ different link functions to relate the probability π to the covariates \mathbf{x} . This family of models was first proposed by Dyke and Patterson (1952), who described a generalization of the ANOVA model for proportions. These models are also members of the family of *generalized linear models (GLMs)* described in Section A.10.

For the i th observation, let π_i be some smooth function of the covariates \mathbf{x}_i with parameter vector $\boldsymbol{\theta} = (\alpha \ \beta_1 \ \cdots \ \beta_p)^T$. In *GLM* notation we assume a binomial error structure with a link function $g(\cdot)$ that relates the probabilities to the covariates such that $g(\pi_i) = \alpha + \mathbf{x}'_i \boldsymbol{\beta}$. Thus, $\pi_i = g^{-1}(\alpha + \mathbf{x}'_i \boldsymbol{\beta})$, where $g^{-1}(\cdot)$ is the inverse function. In logistic regression, $g(\cdot)$ is the logit and $g^{-1}(\cdot)$ is the logistic function (inverse logit). The link function $g(\cdot)$ can be any twice-differentiable one-to-one function.

The binomial likelihood and log likelihood in terms of the $\{\pi_i\}$ are given in (7.3) and (7.5). Using the chain rule, the score equations for the parameters are then obtained as

$$U(\boldsymbol{\theta})_\alpha = \frac{\partial \ell}{\partial \pi} \frac{\partial \pi}{\partial \alpha} \quad \text{and} \quad U(\boldsymbol{\theta})_{\beta_j} = \frac{\partial \ell}{\partial \pi} \frac{\partial \pi}{\partial \beta_j}, \quad (7.28)$$

where $\partial \pi_i / \partial \beta_j = \partial [g^{-1}(\alpha + \mathbf{x}'_i \boldsymbol{\beta})] / \partial \beta_j$, $j = 1, \dots, p$. From (7.5), the resulting score equation for β_j then is of the form

$$U(\boldsymbol{\theta})_{\beta_j} = \sum_i \left(\frac{y_i}{\pi_i} - \frac{1 - y_i}{1 - \pi_i} \right) \frac{\partial \pi_i}{\partial \beta_j} = \sum_i \left(\frac{y_i - \pi_i}{\pi_i(1 - \pi_i)} \right) \frac{\partial \pi_i}{\partial \beta_j}, \quad (7.29)$$

which is a weighted sum of standardized residuals; and likewise for $U(\theta)_\alpha$. These expressions provide the elements of the score vector $\mathbf{U}(\theta)$ that provide the maximum likelihood estimating equations for the parameters. Then the elements of the Hessian are obtained as

$$\begin{aligned}\mathbf{H}(\theta)_\alpha &= \frac{\partial^2 \ell}{\partial \alpha^2} = \frac{\partial}{\partial \alpha} \left(\frac{\partial \ell}{\partial \pi} \frac{\partial \pi}{\partial \alpha} \right) = \frac{\partial \ell}{\partial \pi} \frac{\partial^2 \pi}{\partial \alpha^2} + \frac{\partial^2 \ell}{\partial \pi^2} \left(\frac{\partial \pi}{\partial \alpha} \right)^2 & (7.30) \\ \mathbf{H}(\theta)_{\beta_j} &= \frac{\partial^2 \ell}{\partial \beta_j^2} = \frac{\partial}{\partial \beta_j} \left(\frac{\partial \ell}{\partial \pi} \frac{\partial \pi}{\partial \beta_j} \right) = \frac{\partial \ell}{\partial \pi} \frac{\partial^2 \pi}{\partial \beta_j^2} + \frac{\partial^2 \ell}{\partial \pi^2} \left(\frac{\partial \pi}{\partial \beta_j} \right)^2 \\ \mathbf{H}(\theta)_{\alpha, \beta_j} &= \frac{\partial^2 \ell}{\partial \alpha \partial \beta_j} = \frac{\partial}{\partial \beta_j} \left(\frac{\partial \ell}{\partial \pi} \frac{\partial \pi}{\partial \alpha} \right) = \frac{\partial \ell}{\partial \pi} \frac{\partial^2 \pi}{\partial \alpha \partial \beta_j} + \frac{\partial^2 \ell}{\partial \pi^2} \left(\frac{\partial \pi}{\partial \alpha} \right) \left(\frac{\partial \pi}{\partial \beta_j} \right) \\ \mathbf{H}(\theta)_{\beta_j, \beta_k} &= \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k} \left(\frac{\partial \ell}{\partial \pi} \frac{\partial \pi}{\partial \beta_j} \right) = \frac{\partial \ell}{\partial \pi} \frac{\partial^2 \pi}{\partial \beta_j \partial \beta_k} + \frac{\partial^2 \ell}{\partial \pi^2} \left(\frac{\partial \pi}{\partial \beta_j} \right) \left(\frac{\partial \pi}{\partial \beta_k} \right)\end{aligned}$$

for $1 \leq j < k \leq p$, from which the observed and expected informations are obtained. These expressions are special cases of the more general equations for the family of *GLMs* described in Section A.10.

As a problem, the model equations for an exponential risk model are obtained using the simple log link, and for a compound exponential model using the complementary log-log link. Also included in this family of models is probit regression using the inverse probit link.

These models differ in the manner that the Bernoulli residuals $\{y_i - \pi_i\}$ are weighted in the score estimating equation. From (7.29), the score equation for the j th coefficient β_j can be expressed as

$$U(\theta)_{\beta_j} = \sum_i w_{ij} (y_i - \pi_i), \quad (7.31)$$

with weights

$$w_{ij} = \left(\frac{1}{\pi_i(1 - \pi_i)} \right) \frac{\partial \pi_i}{\partial \beta_j}. \quad (7.32)$$

For example, in a logit link (logistic regression) model, $\partial \pi / \partial \beta_j = x_{ij} \pi_i (1 - \pi_i)$ and the weights are $w_i = x_{ij}$. However, in an exponential risk model with a log link, $\partial \pi / \partial \beta_j = x_{ij} \pi_i$ and the weights are $w_i = x_{ij} / (1 - \pi_i)$.

7.2 INTERPRETATION OF THE LOGISTIC REGRESSION MODEL

7.2.1 Model Coefficients and Odds Ratios

In a logistic regression model, each estimated coefficient equals the estimated log odds ratio associated with a unit increase in the value of the covariate, adjusted for all other effects in the model, or assuming that the values of all other covariates are held fixed. To see this, consider the effect of a single covariate, say the last X_p . The effects of all the other $p - 1$ covariates in the model with fixed values (x_1, \dots, x_{p-1})

are held constant, expressed as $\mathbf{C} = \sum_{j=1}^{p-1} x_j \beta_j$. The effect of the last covariate is $x_p \beta_p$, or simply $x\beta$. Then the model can be parameterized using $\mathbf{x}'\beta = \mathbf{C} + x\beta$.

First consider a binary covariate with only two categories coded as $X = 1$ if exposed (E), 0 if not exposed (\bar{E}). Then the ratio of the odds for an exposed individual to those of an individual not so exposed is

$$OR_{E:\bar{E}} = \frac{O_{x=1}}{O_{x=0}} = \frac{e^{\alpha + \mathbf{C} + \beta}}{e^{\alpha + \mathbf{C}}} = e^\beta. \quad (7.33)$$

Therefore, the log(odds ratio) = β for $x = 1$ versus $x = 0$, or for E versus \bar{E} .

However, the user must be careful because in some cases SAS enters a binary covariate into the design matrix coded as a contrast using values 1 if exposed (E) and -1 if not exposed (\bar{E}), in which case $\exp(2\beta) = OR$ (see Section 7.2.2).

Now consider a quantitative covariate. The odds ratio associated with a 1-unit difference in the value of the covariate is

$$OR_{\Delta x=1} = \frac{O_{x+1}}{O_x} = \frac{e^{\alpha + \mathbf{C} + \beta(x+1)}}{e^{\alpha + \mathbf{C} + \beta x}} = e^\beta. \quad (7.34)$$

Therefore, $[\Delta \log(\text{odds}) | \Delta x = 1] = \beta$. Thus, for a binary covariate coded such that $\Delta x = 1$, and for any quantitative covariate, $\hat{\beta}$ equals the log odds ratio associated with a unit change (increase) in the covariate.

Similarly, the change in the log odds associated with a K -unit difference in the value of the covariate is $[\Delta \log(\text{odds}) | \Delta x = K] = K\beta$ with the associated odds ratio

$$OR_{\Delta x=K} = e^{K\beta} = (OR_{\Delta x=1})^K, \quad (7.35)$$

and estimated standard error

$$S.E.(K\hat{\beta}) = |K| S.E.(\hat{\beta}). \quad (7.36)$$

This also allows one to readily invert the odds ratio, or to describe the odds ratio associated with a K unit *decrease* in X using some $K < 0$. For example, $\beta < 0$ implies that there is a decrease in the odds associated with an increase in X ; however, in many cases it is easier to explain the association in terms of the increase in the odds associated with a decrease in X . For example, if exposure is associated with a decrease in the odds, $OR_{E:\bar{E}} < 1$, then the increase in the odds associated with nonexposure is simply $OR_{\bar{E}:E} = e^{-\beta}$. Likewise, if an increase in a quantitative covariate is associated with a decrease in the odds, then the relationship is readily inverted to describe the increase in odds associated with a decrease in the covariate. In both cases $K = -1$.

For any K , the asymmetric $1 - \alpha$ confidence interval on the odds ratio associated with a K unit change in the covariate is obtained as

$$\exp \left[K\hat{\beta} \pm Z_{1-\alpha/2} |K| \hat{V}(\hat{\beta})^{1/2} \right], \quad (7.37)$$

where $\hat{V}(\hat{\beta})$ is obtained from the corresponding diagonal element of the inverse estimated information matrix.

Table 7.5 SAS program for DCCT nephropathy data.

```

data renal;
input obsn micro24 int hbael duration sbp female;
yearsdm=duration/12; cards;
  1 0 1 9.63 178 104 1
  2 0 0 7.93 175 112 0
  3 1 0 11.20 126 110 1
  4 1 0 10.88 116 106 0
  5 0 0 8.22 168 110 1
  6 1 1 12.73 71 112 0
  7 0 0 8.28 107 116 1
  8 0 1 9.44 79 120 1
  9 0 0 7.44 176 120 1
  10 0 0 8.33 47 140 0
  11 1 1 9.89 135 126 1
  12 0 1 12.06 117 120 0
  13 0 1 9.01 35 128 1
  14 0 1 10.05 82 110 1
  15 1 1 9.60 70 108 1
  16 0 1 10.17 159 121 1
  ...
  157 0 0 9.70 132 106 1
  158 1 0 11.80 157 128 0
  159 0 1 7.00 64 126 0
  160 0 1 9.00 141 114 0
  161 0 1 8.00 46 131 0
  162 1 0 12.50 99 116 1
  163 1 0 7.10 89 114 0
  164 0 0 8.60 139 130 1
  165 0 1 12.20 76 106 1
  166 0 1 11.70 118 110 1
  167 1 0 11.30 99 122 1
  168 0 0 6.80 41 104 1
  169 0 0 10.60 82 106 1
  170 0 1 8.70 101 98 1
  171 0 0 7.90 136 126 0
  172 0 0 10.10 127 124 0
;
proc logistic descending;
model micro24 = int hbael yearsdm sbp female
/ rl covb rsquare;
units int=1 -1 hbael=1 yearsdm=1 sbp=1 5 female=1 -1;
run;

```

Table 7.6 Logistic regression analysis of DCCT nephropathy data.

The LOGISTIC Procedure - Model Fit Statistics						
Criterion	Intercept Only	Intercept and Covariates				
-2 Log L	191.215	155.336				
R-Square	0.1883	Max-rescaled R-Square	0.2806			
Testing Global Null Hypothesis: BETA=0						
Test	Chi-Square		DF	Pr > ChiSq		
Likelihood Ratio	35.8787		5	<.0001		
Score	33.7730		5	<.0001		
Analysis of Maximum Likelihood Estimates						
Parameter	Standard	Wald	Pr >	Odds		
Variable	DF	Estimate	Error	Chi-Square	Chi-Square	
INTERCPT	1	-8.2806	3.0714	7.2686	0.0070	
INT	1	-1.5831	0.4267	13.7626	0.0002	
HBAEL	1	0.5675	0.1449	15.3429	0.0001	
YEARSDM	1	0.00961	0.0636	0.0228	0.8799	
SBP	1	0.0233	0.0208	1.2579	0.2620	
FEMALE	1	-0.8905	0.4473	3.9641	0.0465	
Conditional Odds Ratios and Wald 95% Confidence Intervals						
Variable	Unit	Ratio	Odds	Confidence Limits		
INT	1.0000	0.205	0.089	0.474		
INT	-1.0000	4.870	2.110	11.240		
HBAEL	1.0000	1.764	1.328	2.343		
YEARSDM	1.0000	1.010	0.891	1.144		
SBP	1.0000	1.024	0.983	1.066		
SBP	5.0000	1.123	0.917	1.377		
FEMALE	1.0000	0.410	0.171	0.986		
FEMALE	-1.0000	2.436	1.014	5.854		
Estimated Covariance Matrix						
Variable	INTERCPT	INT	HBAEL	YEARSDM	SBP	FEMALE
INTERCPT	9.43355	-0.0125	-0.2582	-0.0401	-0.0550	-0.1975
INT	-0.0125	0.18211	-0.0097	-0.0005	0.00032	0.02425
HBAEL	-0.2582	-0.0097	0.02100	0.00154	0.00044	-0.0158
YEARSDM	-0.0401	-0.0005	0.00154	0.00405	-0.0001	-0.0023
SBP	-0.0550	0.00032	0.00044	-0.0001	0.00043	0.00243
FEMALE	-0.1975	0.02425	-0.0158	-0.0023	0.00243	0.20005

Similar developments can be used to describe the effects of covariates on the response variable after transformations of either. For example, in a log-log linear model where $\log(Y)$ is regressed on $\log(X)$, then the coefficient β for $\log(X)$ is such that the proportionate change in Y associated with a c -fold change in X is c^β (see Problem 7.5).

Example 7.4 DCCT Nephropathy Data

Section 1.5 provides a description of the Diabetes Control and Complications Trial (DCCT) and presents an analysis of the cumulative incidence of the onset of microalbuminuria that is the earliest sign of diabetic renal disease (early nephropathy). The lifetable analyses of cumulative incidence are described in Chapter 9. Another way to assess the effect of treatment on the risk of nephropathy is to examine the prevalence of microalbuminuria at a fixed point in time. Because the average duration of follow-up was about six years, we present analyses of the factors related to the prevalence of microalbuminuria in those subjects evaluated at six years. Further, we consider the subset of 172 patients in the secondary intervention cohort with a high-normal albumin excretion rate (*AER*) at baseline defined as $15 \leq AER \leq 40$. In addition to intensive versus conventional treatment group (*int* = 1 or 0, respectively), the analysis adjusts for the level (%) of HbA_{1c} at baseline (*hbael*), the prior duration of diabetes in months (*duration*), the level of systolic blood pressure in mmHg (*sbp*), and gender (*female*) (1 if female, 0 if male).

The HbA_{1c} is a measure of the average level of blood glucose control over the preceding 4 to 6 weeks. The measure employed here (*hbael*) is that obtained at the initial eligibility screening that reflects the level of control prior to participation in the trial. The underlying hypothesis is that the level of hyperglycemia and duration of diabetes together determine the risk of further progression of complications of diabetes. Here the months duration is converted to years duration of diabetes (*yearsdm*). Risk of progression of nephropathy may also be associated with increased levels of blood pressure that lead to damage to the kidneys, that also may increase the risk of progression of nephropathy. These effects may also differ for men and women. Thus, the objective is to obtain an adjusted assessment of the effects of intensive versus conventional treatment and also to explore the association between these baseline factors and the risk of nephropathy progression.

The SAS program in Table 7.5 fits a logistic regression model using PROC LOGISTIC. The output from the program is presented in Table 7.6 (extraneous information has been deleted). Note that while *int* and *female* are both qualitative covariates (or *class* variables in SAS), in this model specification both *int* and *female* enter the model directly so that the coefficient reflects the log odds ratio of category 1 versus category 0 (i.e., intensive vs. conventional, and female vs. male). Class effects are described in the next section.

The estimated coefficient for treatment group (*int*) is $\hat{\beta} = -1.5831$ that is the difference in log odds, or the log odds ratio, for intensive versus conventional treatment. The large sample 95% confidence limits on the log odds ratio are obtained as $-1.5831 \pm (1.96)(0.4267) = (-2.4194, -0.7468)$. The estimated odds ratio, therefore, is $\widehat{OR}_{I:C} = e^{-1.5831} = 0.205$ with asymmetric 95% confidence limits of

$(e^{-2.4194}, e^{-0.7468}) = (0.0890, 0.4739)$. The latter values are labeled as the Wald confidence limits because they correspond to confidence limits obtained by inverting the Wald test statistic for the parameter (described in Section 7.3.3). Thus, intensive treatment results in a 79.5% reduction in the odds of having microalbuminuria after six years of treatment compared to conventional therapy.

Because the odds ratio is less than 1 ($\hat{\beta} < 0$), there is a decrease in risk (odds) in the intensive therapy group. The complement of this coefficient provides the increase in odds (odds ratio) for conventional versus intensive therapy, $\widehat{OR}_{C:I} = e^{1.5831} = 4.87$ with asymmetric 95% confidence limits of $(e^{0.7468}, e^{2.4194}) = (2.11, 11.24)$. Thus, conventional therapy yields a 4.87-fold increase in the odds versus intensive therapy.

Similarly, the estimated coefficient for female gender ($\hat{\beta} = -0.8905$) yields an odds ratio of $\widehat{OR}_{F:M} = 0.410$ for females versus males. The inverse yields an estimated odds ratio for males versus females of $\widehat{OR}_{M:F} = 2.436$ with 95% C.I. = (1.014, 5.854).

Hbael is a quantitative covariate measured in percent, since it is the percent of red cells (hemoglobin) that have been glycosylated in reaction with free glucose in systemic circulation. Thus, the estimated coefficient $\hat{\beta} = 0.5675$ is the difference in log odds (the log odds ratio) for each 1% higher HbA_{1c}; such as the difference in the log odds for two subjects, one with a HbA_{1c} of 8 and another 9, or one a value of 10 and another 11. The 95% C.I. is $0.5675 \pm (1.96)(0.1449) = (0.2835, 0.8515)$. Thus, the estimated odds ratio associated with a 1% higher HbA_{1c} is $\widehat{OR} = e^{0.5675} = 1.7639$ with 95% Wald C.I. of $(e^{0.2835}, e^{0.8515}) = (1.328, 2.343)$.

Similarly, the estimated coefficient for years duration ($\hat{\beta} = 0.0096$) implies an odds ratio of 1.010 per year greater duration with 95% C.I. (0.891, 1.144); and the estimated coefficient for systolic blood pressure ($\hat{\beta} = 0.0233$) implies an odds ratio of 1.024 per mmHg higher pressure with 95% C.I. (0.983, 1.066). However, a 1-mmHg difference in blood pressure is a small difference with respect to the distribution of blood pressure values with a mean of 115.6 and S.D. of 11.3. Thus, it is more relevant to describe the odds ratio associated with perhaps a 5-mmHg higher pressure ($K = 5$). The associated log odds ratio is $5\hat{\beta} = 0.1165$ with S.E. = $5(0.0208) = 0.104$ and 95% C.I. of $(-0.087, 0.320)$. The corresponding odds ratio is $\widehat{OR} = e^{0.1165} = 1.123$ with 95% C.I. = $(e^{-0.087}, e^{0.320}) = (0.917, 1.377)$.

Using the expressions in (7.35) and (7.36), it is also possible to invert the relationship for a quantitative covariate. For example, the coefficient for the HbA_{1c} implies a reduction in risk as the HbA_{1c} is reduced. From the coefficient and its S.E., for a 2 percent reduction in HbA_{1c} ($K = -2$), the corresponding coefficient is $(-2)0.5675 = -1.135$ with S.E. = $2(0.1449) = 0.2898$. These yield an odds ratio of 0.321 with 95% C.I. (0.182, 0.567).

PROC LOGISTIC includes an option for a *units* specification, as shown in Table 7.5. The output in Table 7.6 provides the odds ratio associated with a one-unit change for each covariate and that associated with each of the units change specified in the *units* statement.

Table 7.6 also presents the estimated covariance of the coefficient estimates obtained as the inverse of the estimated information matrix. This is generated by the *covb* option. This covariance matrix is useful for computing the variance of an estimated probability and in other computations, such as the robust variance covariance matrix, that is described later.

7.2.2 Class Effects in PROC LOGISTIC

In Example 7.4, the two discrete covariates (*int* and *female*) were coded as binary variables in the data set so that the category with value 0 represented the reference category and 1 the index category of interest. In this case the coefficient estimate equals the log odds ratio of the outcome for category 1 versus 0.

PROC LOGISTIC also provides a *class* statement that will automatically insert design effects and that allows for a polychotomous class variable. However, the default design effect coding is not a binary variable (or set of binary variables), but rather a contrast, referred to as *EFFECT* parameterization of the *CLASS* variable.

For illustration, the following statements would provide a default coding for the class effects *int* and *female* in the above Example 7.4:

```
proc logistic descending;
class int female;
model micro24 = int hbael yearsdm sbp female / rl;
```

The resulting calculations are presented in Table 7.7 (extraneous material has been removed). The model, with the *descending* option, is again modeling the probability of microalbuminuria. The class effect information shows that a contrast effect is used with values +1 and -1. However, the -1 is assigned to the largest value of the class variable (1 in this case) and the +1 to the smaller value (0). Thus, the coefficient represents the difference between category 0 versus 1 rather than category 1 versus 0 as in Table 7.6, and the sign of the coefficient for each effect is reversed from that in Table 7.6.

Further, with the +1 versus -1 effect coding, the coefficient then represents the change in the logit per 2-unit change in the value of the covariate (class effect). Thus, the estimated coefficient equals the log (odds ratio/2) and $\exp(2\hat{\beta})$ equals the odds ratio for category 0 versus 1. In turn, this odds ratio estimate is the reciprocal of that provided in Table 7.6.

These difficulties can be alleviated by adding an additional specification in the *class* statement as follows:

```
proc logistic descending;
class int female / param=ref descending;
model micro24 = int hbael yearsdm sbp female / rl;
```

The *ref* option specifies that a binary design effect is used for each class effect. However, the default effect provides a code of 0 for the highest-numbered category

Table 7.7 CLASS effects in analysis of DCCT data using PROC LOGISTIC.

Probability modeled is micro24=1.

Class Level Information						
				Design		
	Class	Value		Variables		
Parameter	int	0		1		
		1		-1		
Intercept		0		1		
		1		-1		
Analysis of Maximum Likelihood Estimates						
			Standard		Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq	
Intercept	1	-9.5175	3.0547	9.7075	0.0018	
int	0	1	0.2134	13.7626	0.0002	
:						
female	0	1	0.4453	0.2236	3.9641	0.0465
Odds Ratio Estimates						
			Point		95% Wald	
Effect			Estimate		Confidence Limits	
int	0	vs 1	4.870	2.110	11.240	
:						
female	0	vs 1	2.436	1.014	5.854	

(e.g., *int* = 1) versus 1 for the smaller-numbered category (*int* = 0). To ensure that the class effect uses the lowest-numbered category as the reference, the *descending* option is specified in the *class* statement as well. The resulting computations equal those in Table 7.6.

The *class* statement would also apply to a polychotomous covariate. For the stratified analysis of the ulcer clinical trial data in Example 7.3, the variable *stratum* could be assigned the values 1, 2, or 3 to refer to the three strata. Then the following statements will automatically include the 2 *df* stratum effect in the model:

```
proc logistic descending;
class stratum;
model response=stratum group;
```

7.2.3 Partial Regression Coefficients

Any regression model has three main components, as exemplified by the *GLM* family described in Section A.10.

First is the *structural component*, which specifies the nature of the link between the conditional expectation and the covariates, such as

$$E(Y|\mathbf{X}) = \mu, \quad \text{and } g(\mu) = \alpha + \mathbf{X}'\boldsymbol{\beta}, \quad (7.38)$$

so that

$$E(Y|\mathbf{X}) = g^{-1}(\alpha + \mathbf{X}'\boldsymbol{\beta}) \quad (7.39)$$

for some link function $g(\cdot)$ of the linear predictor $\eta = \alpha + \mathbf{X}'\boldsymbol{\beta}$. In ordinary multiple regression, g is the identity link. In logistic regression g is the logit link. The predictor may be nonlinear, and in some cases, such as *generalized additive models* (Hastie and Tibshirani, 1990), it is replaced by some smooth function of the covariates. The estimated model then provides an estimate of the nature of the association under the assumed model. If one assumes that the logit is a linear function of the covariate, then the model provides an "optimal" estimate of the slope of the relationship. However, this does not guarantee that the estimated model is correct. Thus, it is important to assess the adequacy of the model specifications using techniques such as residual analysis or the analysis of value-added plots, in addition to fitting different expressions for the predictor, such as polynomials or logarithms, and so forth, when analyzing data using these methods. Excellent references for these assessments are Pregibon (1981) and Hosmer and Lemeshow (2004), among many. Many of these model diagnostics are available through PROC LOGISTIC. Model diagnostics, however, will not be assessed herein.

Second is the *random component*, which is usually assumed to be of the form

$$y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \perp \mu_i, \quad (7.40)$$

where the random errors ε_i are assumed to be independent of the conditional expectation μ_i and where the errors have some distribution $\varepsilon_i \sim f(\varepsilon)$ with mean zero and a variance that may be some function of the expectation expressed as

$$E(\varepsilon_i) = 0, \quad V(y_i|\mu_i) = V(\varepsilon_i|\mu_i) = \sigma_\varepsilon^2(\mu_i). \quad (7.41)$$

In ordinary multiple regression we assume that the $\{\varepsilon_i\}$ are *i.i.d.* with $E(\varepsilon_i) = 0$ and constant $V(\varepsilon_i) = \sigma_\varepsilon^2$. In logistic regression we assume that the Y_i are independent (but not identically) distributed Bernoulli variables with mean $\mu_i = \pi_i$ so that $E(\varepsilon_i) = 0$ and $V(\varepsilon_i|\pi_i) = \sigma_\varepsilon^2(\pi_i) = \pi_i(1 - \pi_i)$. The random component specification, in effect, determines how each individual observation is weighted in the computation of the estimates of the model parameters.

For example, in a multiple regression model fit using ordinary least squares, constant variance is assumed such that each observation has equal weight in the computation of the sum of squared errors, the objective function to be minimized. However, if homoscedastic errors do not apply, then the model is fit using weighted

least squares, where each observation is weighted inversely to its variance. Similarly, in a *GLM* the score equation (A.252) is a weighted sum of residuals with weights inversely proportional to the $\sigma_\varepsilon^2(\mu_i)$. In addition, the link function also affects the way that each observation is weighted, as shown in Section 7.1.5.

The above two components of regression models are described in every modern text on regression. However, the third component of importance, the *covariate specification*, is rarely discussed. Any model includes a specific set of covariates (X_1, \dots, X_p) , where the elements are selected from a larger set $X_j \in \Omega_\chi$, Ω_χ being the set of possible covariates in the model. If one adds or subtracts a covariate from the model, where a transformation of a measurement is considered a new covariate, then the *meaning* of the coefficients for the other covariates in the model changes, not just their values.

For example, consider an ordinary regression model with two covariates X_1 and X_2 . To be mathematically precise, the structural component of the assumed model should be expressed as

$$E(Y | X_1, X_2) = \alpha_{|\beta_1, \beta_2} + X_1\beta_{1|\alpha, \beta_2} + X_2\beta_{2|\alpha, \beta_1}. \quad (7.42)$$

If X_3 is now added to the model, then the precise model specification is that

$$\begin{aligned} E(Y | X_1, X_2, X_3) = & \alpha_{|\beta_1, \beta_2, \beta_3} + X_1\beta_{1|\alpha, \beta_2, \beta_3} \\ & + X_2\beta_{2|\alpha, \beta_1, \beta_3} + X_3\beta_{3|\alpha, \beta_1, \beta_2}. \end{aligned} \quad (7.43)$$

In general, these sets of coefficients differ, so that

$$\begin{aligned} \alpha_{|\beta_1, \beta_2, \beta_3} & \neq \alpha_{|\beta_1, \beta_2}, \\ \beta_{1|\alpha, \beta_2, \beta_3} & \neq \beta_{1|\alpha, \beta_2}, \\ \beta_{2|\alpha, \beta_1, \beta_3} & \neq \beta_{2|\alpha, \beta_1}. \end{aligned} \quad (7.44)$$

Thus, the value, and more important, the meaning of each coefficient depends explicitly on the other covariates in the model. Each coefficient is a *partial coefficient* adjusted for the other covariates in the model.

In a simple least squares regression model, when all effects in the design matrix \mathbf{X} are jointly orthogonal (uncorrelated) such that $\mathbf{X}'\mathbf{X}$ is a diagonal matrix, then the coefficients for each covariate are unique and are unchanged in all reduced models containing subsets of the original p covariates. This would apply, for example, to the ANOVA model for a balanced multifactorial design. However, if the covariates are correlated, then the coefficients in any reduced model will change from one model to the next (cf. Snedecor and Cochran, 1967, p. 393). In a logistic regression model, however, even when the covariates are uncorrelated, the value and the meaning of model coefficients change when variables are added to or removed from a model because the link $g(\cdot)$, the logit, is a nonlinear function. Thus, the meaning of each coefficient depends explicitly on the set of covariates included in the model.

To see this, consider the case of only two covariates, where X_1 and X_2 are independent. The logistic model specifies that

$$E(Y | X_1, X_2) = g^{-1}(\alpha_{|\beta_1, \beta_2} + X_1\beta_{1|\alpha, \beta_2} + X_2\beta_{2|\alpha, \beta_1}), \quad (7.45)$$

where $g(\cdot)$ is the logit and $g^{-1}(\cdot)$ is the logistic function (inverse logit). If we now drop X_2 from the model, then we obtain the reduced model

$$E(y|X_1) = g^{-1}(\alpha_{|\beta_1} + X_1\beta_{1|\alpha}). \quad (7.46)$$

However, from (7.45) it follows that the reduced model satisfies

$$\begin{aligned} E(y|X_1) &= E_{X_2}[E(Y|X_1, X_2)] \\ &= E_{X_2}[g^{-1}(\alpha_{|\beta_1, \beta_2} + X_1\beta_{1|\alpha, \beta_2} + X_2\beta_{2|\alpha, \beta_1})]. \end{aligned} \quad (7.47)$$

Were it not for the logistic function $g^{-1}(\cdot)$, then $E_{X_2}(X_2\beta_{2|\alpha, \beta_1})$ might be absorbed into the intercept and $\beta_{1|\alpha} = \beta_{1|\alpha, \beta_2}$. However, with the logit link, $\beta_{1|\alpha} \neq \beta_{1|\alpha, \beta_2}$, even when X_1 and X_2 are independent (see also Problem 7.6).

7.2.4 Model Building: Stepwise Procedures

As with any regression procedure, it is possible to build a logistic regression model that meets some specified criteria for optimality or superiority for a given data set. Model building refers to an iterative process by which one explores the sensitivity of the model to each of the three components of the model specification: (1) the structural relationship represented by the expression for the predictor (linear or nonlinear) and the link function; (2) the error distribution and the correlation structure of the errors, if any; and (3) the elements of the covariate vector that are included in the model.

Various diagnostic tools, such as residuals, can be used with logistic regression to assist in the specification of the form of the prediction and link functions and the error distribution. See Pregibon (1981) and the text by Hosmer and Lemeshow (1989). For example, the value-added plots can serve as a visual aid to determine whether a linear or nonlinear covariate effect best fits the data. Hastie and Tibshirani (1990) describe a general approach to determining the structural elements of a model using generalized additive models. In these models, rather than specifying a form for the predictor and the link function, local "nonparametric" smoothing algorithms such as a kernel or spline smoother are applied to the relationship between X and Y to estimate $E(Y|X)$ as some smooth function of X with a possibly irregular shape. However, these models are computer intensive and have the disadvantage that individual coefficients no longer describe the effects of covariates on the risk of the response in direct terms, such as a partial odds ratio.

Of the three model components, covariate selection is the easiest and most commonly explored using some optimization algorithm, such as stepwise model building. In stepwise models, one starts with a *full model* comprising the complete set of, say, p candidate covariates, including transformations of covariates, such as quadratic, and higher-order effects and interactions. The object is to identify a subset of $r \leq p$ covariates that yield a "best-fitting" model according to some criteria. One approach is to fit all $2^p - 1$ sets of one or more covariates and to select the model with perhaps the greatest model likelihood ratio test value, or the smallest model p -value, among many possibilities. This would be computationally tedious. As an approximation to this

approach for normal errors multiple regression models, Efroymson (1960) suggested a *stepwise procedure* starting with forward covariate selection followed interchangeably with backward elimination based on some criterion. The most commonly used is a test of significance of each covariate, adjusting for the other covariates in the model. Most computer programs for regression models afford the option for some variation of stepwise model building, including PROC LOGISTIC in SAS.

In *forward selection* all remaining candidate covariates not in the model are tested to determine that one, if any, that best meets the inclusion criteria, usually a p -value $\leq \alpha_E$, the minimum significance level for entry into the model, termed the *SLE* in SAS. At the initial step with no variables yet added to the model, all p candidate covariates are tested for inclusion, and the most significant is added provided that its p -value is $\leq \alpha_E$. The process continues until no additional covariates meet the criteria for inclusion.

Conversely, in *backward elimination*, one starts with the full model of p candidate covariates and tests each covariate, in turn, to determine that one which has the smallest contribution to the model, such as that with the largest p -value. That covariate is then eliminated from the model provided that its p -value $\geq \alpha_S$, the maximum significance level for staying in the model, termed the *SLS* in SAS. The model is then refit and the process repeated until none of the covariates remaining in the model meet the criteria for elimination. Mantel (1970) suggested that backward elimination was preferred to forward selection because it required fewer models to be fit and because it would be more likely to identify synergistic combinations of covariates that contributed to the model when those same covariates individually (marginally) did not.

More information about such model building procedures is provided by one of the many standard texts on linear regression models. In a normal errors linear model with all p covariates, the variance of the estimates is a function of the mean square error on $N - p - 1$ *df* that also serves as the denominator in the *F*-tests of significance of the coefficients in the model. In this case, the variance of the coefficient estimates is reduced as the number of extraneous covariates is deleted. Also, the variance of the estimate of the predicted values (the \hat{y}_i) is also reduced.

However, a feature of model building that often is not addressed in standard texts is the fact that any inferences about the resulting model have been compromised by the application of any of the above optimization criteria to select the model. For example, any confidence limits, tests of significance, or p -values obtained following a stepwise covariate selection procedure are highly biased and are virtually worthless. See, for example, Freedman (1983), Miller (1984) and Freedman and Pee (1989), among many. Thus, there is a trade-off between the reduction in variance of the estimates and the potential increase in bias of the estimates (and p -values) associated with "stepwise" model building.

In the normal errors linear model, Freedman (1983) presents limiting expressions for the value of the model *F*-test when a one-stage deletion at level α_S is conducted. That is, all covariates not significant at the α_S level in the full model are deleted at once and the model is refit. For example, if one starts with $N = 100$ observations with $p = 50$ covariates and applies a one-stage deletion with $\alpha_S = 0.25$, then

under the null hypothesis wherein all 50 covariates represent noise, on average 12.5 ($= p \times \alpha_S$) covariates will be selected and the resulting model F -statistic value will be about 3.95, which would have a $p \leq 0.0001$ on 12, 87 df . Thus, the data-dependent selection of covariates grossly inflates the test statistics. Unfortunately, there is no simple way to correct for this inflation.

In addition, the estimates of the model coefficients following model selection are biased. Consider the simple linear model with two covariates such that $E(Y | X_1, X_2) = \alpha + X_1\beta_1 + X_2\beta_2$. To simplify matters, assume that $V(\hat{\beta}_1) = V(\hat{\beta}_2) = \sigma_\beta^2$ and that we always select either X_1 or X_2 in the final reduced model, depending on whether $\hat{\beta}_1 >$ or $< \hat{\beta}_2$, respectively. Miller (1984) then presents computations of $E(\hat{\beta}_1 | \text{sel})$ or the expected value of $\hat{\beta}_1$ given that X_1 was selected for inclusion in the reduced model. Even when X_1 and X_2 are uncorrelated, such that $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = 0$, the following are the values of $E(\hat{\beta}_1 | \text{sel})$ as a function of β_2 and σ_β^2 when the true value of $\beta_1 = 1.0$:

σ_β^2	β_2				
	0.0	0.5	1.0	1.5	2.0
0.3	1.01	1.05	1.17	1.35	1.57
0.6	1.10	1.15	1.28	1.46	1.66

(reproduced with permission). Thus, the process of covariate selection induces a positive bias in the expected value of the coefficient in the subset of models in which X_1 is the selected covariate. This bias increases as the true value of the coefficient for the competing covariate X_2 increases, and as the variance of the estimates increases. Miller also presents computations when the coefficient estimates are correlated; however, the correlation, whether positive or negative, has a negligible effect on the magnitude of the bias that is induced.

Therefore, any inferences about the values of the coefficients or the fit of the model should only be based on the full model or on an *a priori*-specified reduced model. If a reduced model is derived through a data-dependent algorithm, then it should only be characterized as a model that may work about as well as the full model as a basis for predictions, with no statement of statistical significance or importance associated with the reduced model so identified.

One technique that is useful to assess the performance of a model is *cross validation*, whereby the fit of a model is assessed by applying it to independent observations not used to fit the model. The simplest method is *split-sample validation*, where the model is fit to N_A random observations and then applied to another N_B random observations, designated as the A and B samples, respectively. Often the A and B samples are the split-halves or some other fraction of the total sample. Such split-sample validation allows one to assess the properties of a model from observations that are independent of those used to derive the model. It is often sobering to first conduct a stepwise model building exercise from two independent samples and to find that the reduced models so derived from the A and B samples are disjoint with no variables in common, even though all the selected variables may have been "nominally" significant at $p \leq 0.05$ in the reduced models.

The Coronary Drug Project Research Group (1974) presented an example in which separate reduced models of predictors of heart disease were constructed in split samples. In each sample, the model used the first 10 of 40 potential covariates that were chosen by forward selection, all with $|Z| \geq 2.84$. Thus, the "best" 10-covariate reduced models were selected separately and independently in sample *A* and in sample *B*. However, only two of the selected covariates in both sets of 10 were the same. That is, the other eight covariates in the sample *A* model differed from those in the sample *B* model. This suggests that the other 16 selected covariates could merely be noise. Nevertheless, despite the discrepancy in the covariates selected in each reduced model, each model showed equal predictive power when applied to another independent sample of observations. This illustrates that, in general, there is no true "best model" for any population or set of observations. Rather, one can usually identify a group of reduced models containing different collections of covariates, where each model performs equally well according to some criterion.

There is a large literature on indices of the adequacy of reduced models, such as Mallow's C_p (Mallows, 1973) and Akaike's information criterion (Akaike, 1973) that assess in some sense the bias introduced in reduced models. However, these are global measures that do not help assess the actual level of significance associated with the reduced model or the selected covariates, or to assess the true importance of any individual covariates selected or not selected. There is also a growing literature on cross validation techniques to assess model fit by eliminating k observations from the data set, fitting the model from the $N - k$ remaining observations and then applying the model to the k omitted observations. This is also termed the "leave one out" method ($k = 1$) or manyfold cross validation ($k > 1$). An overall assessment of model fit is then obtained by averaging or aggregating over the $\binom{N}{k}$ possible combinations. Thall et al. (1992) and Thall et al. (1997) described such many-fold cross validation and data splitting as a basis for reduced model selection to ensure that selected covariates are predictive when applied to independent observations. Others, such as Breiman (1992), use the bootstrap and related ideas to validate selected models. Such techniques allow one to differentiate true signal from mere noise in the data; however, they can be very computer intensive.

7.2.5 Disproportionate Sampling

Truett et al. (1967) showed that a logistic model could be used in conjunction with a linear discriminant function to provide an estimate of the posterior probability $P(D|\mathbf{X})$ of membership in the index population of subjects with the positive response or disease given the covariate vector \mathbf{X} . This posterior probability depends explicitly on the prior probability of the response or the prevalence of the disease in the population. In a general random sample from the population, the sample fraction with the disease provides an unbiased estimate of this prior prevalence; that is, $p_1 = m_1/N$ and $E(p_1) = P(D)$. In some cases, however, one does not have a general random sample but, rather, two separate disproportionate samples of m_1 of those with the response (D) and m_2 of those without (\bar{D}). In this case the score equa-

tion for the intercept (7.8) requires that $p_1 = \widehat{\pi} = \sum_i \widehat{\pi}_i / N$, where $E(p_1) \neq P(D)$. Therefore, the estimated posterior probabilities obtained from the estimated model coefficients are biased.

Anderson (1972), however, shows that unbiased estimates of the posterior probabilities are readily obtained by subtracting a constant from the intercept that yields a model of the form

$$\log \left[\frac{\widehat{\pi}_i}{1 - \widehat{\pi}_i} \right] = \mathbf{x}_i' \widehat{\boldsymbol{\theta}} - \log \left(\frac{P(\overline{D})}{P(D)} \right). \quad (7.48)$$

When applied to the linear predictor in (7.15) this also yields unbiased estimates of the confidence limits for the posterior probabilities. This modification of the model, however, has no effect on the estimates of the coefficients, the estimated information matrix, or any of the other characteristics of the model.

7.2.6 Unmatched Case-Control Study

Section 5.1.1 showed that the retrospective odds ratio of exposure given disease derived from the unmatched, retrospectively-sampled case-control 2×2 table also provides an estimate of the population prospective odds ratio of disease given exposure. This suggests that logistic regression can be applied to the analysis of such studies to estimate the covariate effects on the prospective odds ratios. Although cases (D) and controls (\overline{D}) are almost always disproportionately sampled in such studies, nevertheless, from the preceding results, allowance for the true prevalence of cases and controls serves only to add a constant to the intercept or linear predictor, affecting nothing else, (see also Breslow and Day, 1980).

Example 7.5 Ischemic Heart Disease

Dick and Stone (1973) present an unmatched case-control study comparing 146 men with ischemic heart disease (the cases) to 283 controls selected from the general population of males within the same age range as the cases (30–69 years). The study assessed the influence of hyperlipidemia (HL) versus normal lipid levels, smoking at least 15 cigarettes/day (SM) versus not, and hypertension (HT , diastolic blood pressure ≥ 95 mmHg) versus not. Each covariate is coded as a binary indicator variable: 1 if yes, 0 if no. The data are

HL	SM	HT	# Cases	# Controls
0	0	0	15	82
0	0	1	10	37
0	1	0	39	81
0	1	1	23	28
1	0	0	18	16
1	0	1	7	12
1	1	0	19	19
1	1	1	15	8

(reproduced with permission). The logistic regression model estimates of the parameter, *S.E.*, odds ratio, and 95% confidence limits (OR_L and OR_U) for each effect are

Effect	$\hat{\beta}$	<i>S.E.</i>	OR	OR_L	OR_U
Intercept	-1.5312	0.2048			
Hyperlipidemia (<i>HL</i>)	1.0625	0.2316	2.894	1.838	4.556
Smoking (<i>SM</i>)	0.7873	0.2185	2.198	1.432	3.372
Hypertension (<i>HT</i>)	0.3262	0.2242	1.386	0.893	2.150

In such a model, the intercept has no direct prospective interpretation. Each of the estimated coefficients, however, provides an estimate of the log retrospective and the log prospective odds ratio for each risk factor adjusting for the other risk factors in the model. Hyperlipidemia has a nearly threefold greater odds of ischemic heart disease than nonhyperlipidemia, and smoking has over a twofold greater odds than nonsmoking. Hypertension produces a modest increase in the odds.

7.3 TESTS OF SIGNIFICANCE

7.3.1 Likelihood Ratio Tests

7.3.1.1 Model Test In most applications there is no reason to conduct a test of significance of the value of the intercept α , that in effect, describes the overall mean or background odds against which the effects of covariates are assessed. Rather, we wish to conduct a global test of the vector of model covariate coefficients $H_0: \beta = \mathbf{0}$ against $H_1: \beta \neq \mathbf{0}$ for $\beta = (\beta_1 \ \cdots \ \beta_p)^T$ given α in the model. Expressed in terms of the likelihood function, the null hypothesis specifies $H_0: L = L(\alpha)$ and the alternative $H_1: L = L(\alpha, \beta)$. Therefore, as shown in Section A.7.2, the likelihood ratio test is

$$X_{\beta}^2 = -2 \log \left(\frac{L(\hat{\alpha})}{L(\hat{\alpha}, \hat{\beta})} \right) = -2 \log L(\hat{\alpha}) - \left[-2 \log L(\hat{\alpha}, \hat{\beta}) \right]. \quad (7.49)$$

Under H_0 , $X_{\beta}^2 \sim \chi^2$ on p *df*. This is the termed the *likelihood ratio model chi-square test*.

For example, in ordinary normal errors multiple regression, from (A.138) it is readily shown that $-2 \log L(\hat{\alpha})$ is proportional to the total SS(Y), whereas $-2 \log L(\hat{\alpha}, \hat{\beta})$ is proportional to the error SS. Therefore, a model likelihood ratio X_{β}^2 test in a normal errors regression model is proportional to the regression SS.

7.3.1.2 Test of Model Components Often it is of interest to test a hypothesis concerning specific components of the coefficient vector. In this case the coefficient vector for the *full model* of p coefficients is partitioned into two subsets of r and s coefficients ($p = r + s$) such as $\beta = (\beta_r // \beta_s)$ and we wish to test $H_0: \beta_r = \mathbf{0}$.

Again, it should be emphasized that this hypothesis should be formally expressed as $H_0: \beta_{r|s} = \mathbf{0}$, that is, a hypothesis with respect to the partial contribution of β_r when added to a model that contains the subset β_s , or dropped from the model with the full vector β . In this case, the null and alternative hypotheses $H_{0r}: \beta_{r|s} = \mathbf{0}$ and $H_{1r}: \beta_{r|s} \neq \mathbf{0}$ imply

$$\begin{aligned} H_{0r}: L &= L(\alpha, \beta_s) \\ H_{1r}: L &= L(\alpha, \beta_{s|r}, \beta_{r|s}) = L(\alpha, \beta). \end{aligned} \quad (7.50)$$

Then as shown in Section A.7.2.2, the likelihood ratio test is

$$\begin{aligned} X_{\beta_{r|s}}^2 &= -2 \log \left(\frac{L(\hat{\alpha}, \hat{\beta}_s)}{L(\hat{\alpha}, \hat{\beta}_{s|r}, \hat{\beta}_{r|s})} \right) \\ &= \left[-2 \log L(\hat{\alpha}, \hat{\beta}_s) \right] - \left[-2 \log L(\hat{\alpha}, \hat{\beta}) \right], \end{aligned} \quad (7.51)$$

which equals the change in $-2 \log(L)$ when β_r is added to, or dropped from, the model. Under the reduced model null hypothesis, $X_{\beta_{r|s}}^2 \sim \chi^2$ on r df.

To compute this statistic, both the full and reduced models must be fit. The test can then be obtained as the difference between the likelihood ratio model chi-square statistics $X_{\beta_{r|s}}^2 = X_{\beta}^2 - X_{\beta_s}^2$, the $-2 \log L(\hat{\alpha})$ canceling from the expressions for each model X^2 to yield (7.51).

Likewise, a likelihood ratio test can be computed for an individual component of the parameter vector by fitting the model with and then without that variable included in the model. To conduct this test for each component of the model requires using PROC LOGISTIC to fit p submodels, with each variable, in turn, eliminated from the full model. However, the SAS procedure GENMOD will compute these likelihood ratio tests directly through its *type3* option., (see Section 7.3.4).

7.3.2 Efficient Scores Test

7.3.2.1 Model Test As shown in Section A.7.3, the efficient scores test for the model null hypothesis $H_0: \beta = \mathbf{0}$ is based on the score vector $\mathbf{U}(\theta)$ and the expected information $\mathbf{I}(\theta)$, each estimated under the tested hypothesis, designated as $\mathbf{U}(\hat{\theta}_0)$ and $\mathbf{I}(\hat{\theta}_0)$, respectively. Since $\theta = (\alpha // \beta)$, and the hypothesis specifies that $\beta = \mathbf{0}$, then the MLE estimated under this hypothesis is the vector $\hat{\theta}_0 = (\hat{\alpha}_0 // \mathbf{0})$. Thus, we need only obtain the MLE of the intercept under this hypothesis, designated as $\hat{\alpha}_0$.

Again, let $m_1 = \#(y_i = 1) = \sum_i y_i$ and $m_2 = \#(y_i = 0) = \sum_i (1 - y_i) = N - m_1$. Evaluating (7.8) under $H_0: \beta = \mathbf{0}$ yields

$$[\mathbf{U}(\theta)_\alpha]_{\beta=\mathbf{0}} = m_1 - \sum_{i=1}^N \frac{e^\alpha}{1 + e^\alpha} = m_1 - N\bar{\pi} = 0, \quad (7.52)$$

such that

$$\hat{\pi} = \frac{m_1}{N}, \quad (7.53)$$

and

$$\hat{\alpha}_0 = \log \left(\frac{\hat{\pi}}{1 - \hat{\pi}} \right) = \log \left(\frac{m_1}{m_2} \right). \quad (7.54)$$

By definition, the resulting score for the intercept evaluated at the *MLE* is

$$U(\hat{\theta}_0)_\alpha = [U(\theta)_\alpha]_{|\hat{\alpha}_0, \beta=0} = 0. \quad (7.55)$$

Given the estimate $\hat{\alpha}_0$, the score equation for each coefficient as in (7.9) is then evaluated under the tested hypothesis. The score for the j th coefficient is then

$$\begin{aligned} U(\hat{\theta}_0)_{\beta_j} &= [U(\theta)_{\beta_j}]_{|\hat{\alpha}_0, \beta=0} = \sum_i x_{ij} \left[y_i - \frac{e^{\hat{\alpha}_0}}{1 + e^{\hat{\alpha}_0}} \right] \\ &= \sum_i x_{ij} \left(y_i - \hat{\pi} \right) = \sum_i x_{ij} \left(y_i - \frac{m_1}{N} \right) = m_1 \left[\bar{x}_{j(1)} - \bar{x}_j \right], \end{aligned} \quad (7.56)$$

where $\bar{x}_{j(1)}$ is the mean of X_j among the m_1 subjects with a positive response ($y_j = 1$) and \bar{x}_j is the mean in the total sample of N observations. Thus, the score for the j th coefficient is proportional to the observed mean of the covariate among those with the response minus that expected under the null hypothesis. If the covariate is associated with the risk of the response then we expect that $\bar{x}_{j(1)} \neq \bar{x}_j$. For example, if the risk of the response increases as age increases, we would expect the mean age among those with the response to be greater than the mean age in the total population. On the other hand, if there is no association we expect $\bar{x}_{j(1)}$ to be approximately equal to \bar{x}_j . Therefore, the total score vector evaluated under H_0 is

$$\begin{aligned} \mathbf{U}(\hat{\theta}_0) &= m_1 [0 \ (\bar{x}_{1(1)} - \bar{x}_1) \ \cdots \ (\bar{x}_{p(1)} - \bar{x}_p)]^T \\ &= m_1 \left[0 \parallel (\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}})^T \right]^T. \end{aligned} \quad (7.57)$$

The information matrix estimated under H_0 , $\mathbf{I}(\hat{\theta}_0) = \mathbf{I}(\theta)_{|\hat{\alpha}_0, \beta=0}$, then has elements obtained from (7.11)–(7.14) that are

$$\begin{aligned} \mathbf{I}(\hat{\theta}_0)_\alpha &= \sum_i \frac{e^\alpha}{(1 + e^\alpha)^2} = N\bar{\pi}(1 - \bar{\pi}), \\ \mathbf{I}(\hat{\theta}_0)_{\alpha, \beta_j} &= \sum_i x_{ij} \frac{e^\alpha}{(1 + e^\alpha)^2} = N\bar{x}_j \bar{\pi}(1 - \bar{\pi}), \\ \mathbf{I}(\hat{\theta}_0)_{\beta_j, \beta_k} &= \bar{\pi}(1 - \bar{\pi}) \sum_i x_{ij} x_{ik}, \\ \mathbf{I}(\hat{\theta}_0)_{\beta_j} &= \bar{\pi}(1 - \bar{\pi}) \sum_i x_{ij}^2. \end{aligned} \quad (7.58)$$

Therefore, the estimated information matrix under H_0 is

$$\mathbf{I}(\hat{\theta}_0) = \hat{\pi}(1 - \hat{\pi})(\mathbf{X}' \mathbf{X}), \quad (7.59)$$

where \mathbf{X} is the $N \times (p + 1)$ design matrix, the first column being the unit vector.

For given $\boldsymbol{\theta}_0$, since $\mathbf{U}(\widehat{\boldsymbol{\theta}}_0) \stackrel{d}{\approx} \mathcal{N}[0, \mathbf{I}(\boldsymbol{\theta}_0)]$, then the model score test is the quadratic form

$$\begin{aligned} X_p^2 &= \mathbf{U}(\widehat{\boldsymbol{\theta}}_0)' \mathbf{I}(\widehat{\boldsymbol{\theta}}_0)^{-1} \mathbf{U}(\widehat{\boldsymbol{\theta}}_0) \\ &= m_1^2 [0 \parallel (\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}})^T] \frac{(\mathbf{X}' \mathbf{X})^{-1}}{\widehat{\pi}(1 - \widehat{\pi})} [0 \parallel (\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}})^T]^T, \end{aligned} \quad (7.60)$$

that is asymptotically distributed as chi-square on p df under the model null hypothesis $H_0: \boldsymbol{\beta} = \mathbf{0}$.

Asymptotically, a score test is fully efficient, as shown in Section A.7.3. Computationally, the test can be obtained without actually fitting the model, that is, without obtaining the MLEs of the parameter vector and the associated estimated information. Thus, the score test can be computed for a degenerate model or one for which the Newton-Raphson iteration does not converge. However, when a convergent solution exists, the likelihood ratio test, in general, is preferred. When the model is degenerate, then it implies that either there is a linear dependency among the covariates, or that a covariate has nearly constant values among those with a positive or negative response. In such cases the model should be refit after the suspect covariate(s) have been removed, in which case the score test from the degenerate model becomes irrelevant.

7.3.2.2 Test of Model Components Similarly, it would be possible to conduct a score test of specific components of the coefficient vector. However, the score test in this instance offers few advantages over the likelihood ratio test, especially considering that the latter is readily obtained by fitting the full and reduced models of interest. Likewise, score tests of the individual components of the parameter vector are rarely used, because this would require evaluation of the score vector and the information matrix under the null hypothesis for each covariate, in turn.

7.3.3 Wald Tests

As described in Section A.7.1, the Wald model test is obtained as a quadratic form in the parameter estimates and the estimated covariance matrix obtained from the inverse of the estimated information; see (A.148). Although the Wald test is readily computed, the likelihood ratio test, in general, is preferred.

Similarly, a Wald test can be computed for the parameters of a submodel or for individual coefficients in the model. To test the hypothesis $H_0: \beta_j = 0$ for the j th coefficient, the Wald test is readily computed as

$$X_j^2 = \frac{\widehat{\beta}_j^2}{\widehat{V}(\widehat{\beta}_j)}, \quad (7.61)$$

where $\widehat{V}(\widehat{\beta}_j) = [\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1}]_{\beta_j}$ is the corresponding diagonal element of the inverse information matrix. Since Wald tests are based on the variance of the coefficient estimate obtained under the alternative hypothesis, then the significance of a Wald

test can be ascertained from the 95% confidence limits. Thus, these limits are also referred to as Wald confidence limits.

For example, for the case-control study in Example 7.5, the 95% confidence limits for the odds ratio associated with hyperlipidemia and with smoking each do not include 1.0, and thus the corresponding Wald test for each are significant at the 0.05 significance level (two-sided). However, that for hypertension does include 1.0 and thus the Wald test is not significant.

Hauck and Donner (1977) and Vaeth (1985) have shown that the Wald test may be highly anomalous in special, albeit unrealistic cases. In particular, for fixed values of all the coefficients except the last ($\beta_1, \dots, \beta_{p-1}$), as the value of β_p increases away from the null, the Wald test may, in fact, decrease, approaching zero. Mantel (1987) shows that the problem is related to the fact that the Wald test employs the variance estimated under the model (i.e., under the alternative) rather than under the null hypothesis. Thus, the score or likelihood ratio tests of the individual parameters or for a submodel are also preferred over the Wald test.

Example 7.6 DCCT Nephropathy Data (continued)

Table 7.6 of Example 7.4 presents the likelihood ratio and score tests for the overall model. The likelihood ratio test is the difference between the null (intercept only) and full model values of $-2 \log(L)$, or $191.215 - 155.336 = 35.879$ on 5 df , $p < 0.0001$.

The model score test is based on the score vector and inverse information matrix evaluated under the null hypothesis. For these data, the score vector evaluated from (7.57) is

$$U(\hat{\theta}_0) = U[\hat{\alpha}, \beta]_{|\beta=0} = (0 \ -10.73 \ 30.01 \ -7.133 \ 78.33 \ -4.047)^T$$

and the estimated information evaluated under the null hypothesis from (7.59) is $I(\hat{\theta}_0) =$

31.744186	16.425771	294.00468	299.33968	3692.6609	14.395619
16.425771	16.425771	151.85902	154.53759	1907.235	7.3823688
294.00468	151.85902	2791.717	2753.4845	34128.414	138.08536
299.33968	154.53759	2753.4845	3174.566	34875.341	138.92695
3692.6609	1907.235	34128.414	34875.341	433238.01	1623.9366
14.395619	7.3823688	138.08536	138.92695	1623.9366	14.395619

The resulting model score test from (7.60) is 33.773 on 5 df with $p < 0.0001$.

Table 7.6 presents the Wald tests for each coefficient in the model. However, these tests employ the estimated variance of the coefficient estimates obtained under the general alternative hypothesis rather than under the covariate coefficient-specific null hypothesis. Thus, likelihood ratio tests of the coefficients are preferred to these Wald tests. However, PROC LOGISTIC does not compute these likelihood ratio tests for the individual coefficients. Rather, when using PROC LOGISTIC it is necessary to compute these by hand by successively fitting a model without each covariate, in turn, which is compared to the full model with all covariates. For example, to compute the likelihood ratio test for the covariate *int*, the model without *int* is fit yielding a

model likelihood ratio test $X^2 = 20.298$ on 4 df (not shown). Compared to the full model likelihood ratio test in Table 7.6, the resulting likelihood ratio test for *int* is $35.88 - 20.30 = 15.58$ with $p < 0.0001$ on 1 df .

Alternatively, PROC GENMOD can be used to fit the model and to directly compute the likelihood ratio test for each covariate.

7.3.4 SAS PROC GENMOD

The SAS procedure GENMOD can be used to fit various models that are members of the family of generalized linear models (*GLMs*) described in Section A.10. This includes the logistic regression model and related models described in Section 7.1.5. In this formulation, a logistic regression model uses the logit link in conjunction with a binomial variance that equals the Bernoulli variance since each observation is a separate individual. PROC GENMOD also provides additional computations that augment those provided by PROC LOGISTIC.

For models with categorical covariates, as in PROC LOGISTIC, GENMOD allows a *class* statement that automatically generates the appropriate number of binary reference contrasts in the design matrix. The *lsmeans* statement then provides covariate-adjusted estimated logits for any class (categorical) effects in the model. LSMEANS stands for *least squares means* as computed in PROC GLM for a normal errors linear model with design (class) effects and quantitative covariate effects. Consider a logistic regression model of the form

$$\text{logit}(\pi_i) = \alpha + X_{1i}\beta_1 + X_{2i}\beta_2 \quad (7.62)$$

where $X_1 = 1$ if treated, 0 if control and X_2 is an adjusting covariate (or a vector of adjusting covariates). For the N observations with complete values (y_i, x_{1i}, x_{2i}) that enter into the model computations, let \bar{x}_2 denote the mean of X_2 among the N observations. Then the LSMEAN adjusted logits for the two treatment groups are

$$\begin{aligned} \text{Treated } (X_1 = 1): \quad \text{logit}(\hat{\pi}_1) &= \hat{\alpha} + \hat{\beta}_1 + \bar{x}_2 \hat{\beta}_2 \\ \text{Control } (X_1 = 0): \quad \text{logit}(\hat{\pi}_0) &= \hat{\alpha} + \bar{x}_2 \hat{\beta}_2 \end{aligned} \quad (7.63)$$

This generalizes to adjustment for multiple covariates, in which case the covariate mean for each covariate is employed. For categorical covariates, the mean of the corresponding binary effects (i.e., the sample proportions) would be employed. Since each LSMEAN logit is a linear combination of the vector of model coefficient estimates, then the variance of the LSMEAN logit is obtained from the corresponding quadratic form. This also provides for the computation of confidence intervals on the LSMEAN logits. Then the adjusted LSMEAN estimated probabilities, and their confidence limits, are obtained from the inverse (logistic) function of the corresponding logits. This can be computed using additional statements, as shown in the following example.

GENMOD also provides a *type3* option for computation of the likelihood ratio tests for the individual coefficients in the model, called Type III tests of covariate

effects. This terminology is used to refer to the test of the partial effect of a covariate in the SAS PROC GLM that fits normal errors regression models.

In *GLM* terminology, the deviance equals the difference in $-2 \log(L)$ for the present model versus a model that fits the data perfectly (see Section A.10.3). For a logistic regression model with a binary response, but not binomial regression, the saturated model log likelihood is zero, so that in this special case, *deviance* = $-2 \log(L)$. The scaled deviance and Pearson chi-square are also described in the Appendix.

Example 7.7 DCCT Nephropathy Data (continued)

The following SAS code fits a logistic regression model to the DCCT data using PROC GENMOD with the LSMEANS for the treatment group (*int*) effect and the type III likelihood ratio tests:

```
proc genmod data = renal descending;
class int female / descending;
model micro24 = int hbael yearsdm sbp female
  / dist=binomial link=logit type3 aggregate;
lsmeans int / cl;
```

In PROC GENMOD, the *descending* option can be used in the *class* statement to define the lowest-ordered category as the reference for computing coefficients and odds ratios, such as the category zero for a binary covariate. The resulting output from the fitted model is presented in Table 7.8 (extraneous information deleted).

In Table 7.8 the *aggregate* option provides the computation of the deviance and Pearson chi-square measures of goodness of fit. These are described in the following section. The *class* statement provides estimates of the regression coefficients for each category versus the reference category, zero in these instances. The estimates and Wald tests are followed by the likelihood ratio (*type3*) tests for each coefficient. Although the Wald test is occasionally greater than the likelihood ratio test, in general the latter has a type I error probability that is closer to the desired level and is more efficient.

The *lsmeans* statement with the *cl* option specified computation of the LSMEAN logit for each category of the treatment group variable (*int*) as described in (7.63). This is computed as a linear function of the coefficient estimates $L'\beta$, where β is the complete vector of model parameter estimates, including the intercept, and L' is a vector consisting of the covariate value corresponding to the design effect (*int* = 0 or 1 in this case) and the covariate means for all other effects. For these data

	Intercept	INT	hbael	yearsdm	sbp	female
L' =	1	0 or 1	9.2617	9.4297	116.33	0.4535
β =	-8.2806	-1.5831	0.5675	0.0096	0.0233	-0.8905

Thus, the adjusted $\text{logit}(\hat{\pi}_i) = L'_i \beta$, where $i = 0, 1$, corresponding to the two treatment groups. The variance of the logit is obtained as $L'_i \hat{\Sigma}_\beta L_i$, where $\hat{\Sigma}_\beta$ is the estimated covariance matrix of the coefficient estimates. These provide the terms "L'Beta" and "Standard Error" in the LSMEANS output in Table 7.8. The *cl* option

Table 7.8 Logistic regression analysis of DCCT data using PROC GENMOD.

The GENMOD Procedure			
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	166	155.3362	0.9358
Scaled Deviance	166	155.3362	0.9358
Pearson Chi-Square	166	182.2750	1.0980
Scaled Pearson X2	166	182.2750	1.0980
Log Likelihood	.	-77.6681	.

**Analysis Of Maximum Likelihood
Parameter Estimates**

Parameter	DF	Estimate	Std Err	Wald	
				Chi-Square	Pr>Chi
INTERCEPT	1	-8.2806	3.0731	7.2686	0.0005
INT	1	-1.5831	0.4267	13.7626	0.0002
HBAEL	1	0.5675	0.1449	15.3429	0.0001
YEARSDM	1	0.0096	0.0636	0.0228	0.8799
SBP	1	0.0233	0.0208	1.2579	0.2620
FEMALE	1	-0.8905	0.4473	3.9641	0.0465
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
INT	1	15.5803	0.0001
HBAEL	1	17.6701	0.0001
YEARSDM	1	0.0229	0.8798
SBP	1	1.2726	0.2593
FEMALE	1	4.1625	0.0413

Least Squares Means

Effect	INT	Mean	L'Beta	Standard Error	Confidence Limits	
					Estimate	Standard
INT	1	0.0951	-2.2532	0.3638	-2.9662	-1.5403
INT	0	0.3385	-0.6701	0.2545	-1.1689	-0.1713

then provides the computation of the large sample confidence limits on the adjusted logit.

Taking the inverse (logistic) function yields the covariate-adjusted estimated probability within each group labeled "Mean" in the output. Thus, $\hat{\pi}_1 = [1 + \exp(-L'_1\beta)]^{-1} = [1 + \exp(2.2532)]^{-1} = 0.0951$ and $\hat{\pi}_0 = [1 + \exp(0.6701)]^{-1} = 0.3385$. Applying the logistic function to the confidence limits on $L'\beta$ then yields confidence limits on these adjusted probability estimates of (0.04873, 0.1773) in the intensive group and (0.2364, 0.4582) in the conventional group. Also, note that the odds ratio provided by these adjusted probabilities equals the value 0.205 provided by the model in Table 7.6.

These computations are provided by additional statements following the call of the procedure as shown in the program *renal.sas*.

7.3.5 Robust Inferences

All regression models that are based on a full likelihood specification, such as the *GLM* family in general, and the logistic regression model in particular, explicitly assume that the variance-covariance structure specified by the model is correct and applies to the population from which the observations were obtained. In the logistic model, conditional on the covariate vector value \mathbf{x} , the assumption is that $V(y|\mathbf{x}) = \pi(\mathbf{x})[1 - \pi(\mathbf{x})]$, where $E(y|\mathbf{x}) = \pi(\mathbf{x})$ is the conditional expectation (probability). In some cases, however, the variation in the data may be greater or less than that assumed by the model; that is, there may be over- or underdispersion. In this case, the properties of confidence limits and test statistics can be improved by using an estimate of the variance-covariance matrix of the estimates that is robust to departures from the model variance assumptions.

One simple approach is to fit an over (under)-dispersed *GLM* regression model by adopting a quasi-likelihood with an additional dispersion parameter (see Section A.10.4). This approach is widely used in Poisson regression models of count data and thus is described in Chapter 8. However, it could also be employed with logistic regression. As described in Section A.10.3, the ratio of the Pearson chi-square test value to its degrees of freedom is an estimate of the degree of departure from the model-specified variance assumption, a value substantially greater than one indicating overdispersion, substantially less indicating underdispersion. When overdispersion exists, the model underestimates the degree of variation in the data, and thus the variance-covariance matrix of the coefficients is underestimated, so that Wald tests are inflated and confidence intervals are too narrow. The opposite occurs with underdispersion.

For example, in the analysis of the DCCT nephropathy data in Table 7.8, $\chi^2/\text{df} = 1.098$, which suggests that the actual variance-covariance matrix of the estimates may be about 10% greater than that estimated from the model based on the inverse estimated information matrix. This degree of overdispersion is well within the range one might expect under random variation with 95% confidence when the model assumptions actually apply, computed as $1 \pm [2.77/\sqrt{\text{df}}] = 2.77/\sqrt{166} = 0.215$ (see Section A.10.3). However, if substantial departures are suggested, and there are

no obvious deficiencies or errors in the model specification, then this approach can be used to estimate the overdispersion scale parameter and to adjust the estimates of the variances of the model coefficients. Section 8.3.2 describes the application of this approach to Poisson regression using PROC GENMOD.

Another approach to adjust for departures from the model variance assumption is to employ the *robust information sandwich estimator* described in Section A.9. The information sandwich provides a consistent estimate of the variance of the estimates for any model where the first moment specification (the structural component) is correct, but where the second moment specification (the error variance-covariance structure) may not be correctly specified.

Let \mathbf{X} denote the $n \times (p + 1)$ design matrix, where the i th row is the covariate vector \mathbf{x}_i^T for the i th subject augmented by the constant (unity) for the intercept. Also, let $\boldsymbol{\Gamma} = \text{diag}[\pi_i(1 - \pi_i)]$, where $\pi_i = E(y_i|\mathbf{x}_i)$ is the conditional probability expressed as a logistic function of the covariates. Then, from (7.11)–(7.14), the expected information can be expressed as

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{X}'\boldsymbol{\Gamma}\mathbf{X}. \quad (7.64)$$

This is the covariance matrix of the score vector when the specified logistic model is assumed to be correct.

However, when the model-specified covariance matrix is not correct, let $\boldsymbol{\Sigma}_\varepsilon = \text{diag}[E(y_i - \pi_i)^2]$ refer to the true covariance matrix of the errors. Then, from (7.8)–(7.9), it follows that the true covariance matrix of the score vector is

$$\mathbf{J}(\boldsymbol{\theta}) = \sum_i E[\mathbf{U}_i(\boldsymbol{\theta})\mathbf{U}_i(\boldsymbol{\theta})'] = \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon\mathbf{X}. \quad (7.65)$$

From (A.229), the expression for the robust information sandwich covariance matrix of the coefficient estimates is then

$$\boldsymbol{\Sigma}_R(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\boldsymbol{\theta})^{-1}\mathbf{J}(\boldsymbol{\theta})\mathbf{I}(\boldsymbol{\theta})^{-1} = (\mathbf{X}'\boldsymbol{\Gamma}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Sigma}_\varepsilon\mathbf{X})(\mathbf{X}'\boldsymbol{\Gamma}\mathbf{X})^{-1}. \quad (7.66)$$

This matrix can be consistently estimated as

$$\hat{\boldsymbol{\Sigma}}_R(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}, \quad (7.67)$$

where $\mathbf{I}(\hat{\boldsymbol{\theta}}) = \mathbf{X}'\hat{\boldsymbol{\Gamma}}\mathbf{X}$, $\hat{\boldsymbol{\Gamma}} = \text{diag}[\hat{\pi}_i(1 - \hat{\pi}_i)]$, $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}) = \sum_i [\mathbf{U}_i(\hat{\boldsymbol{\theta}})\mathbf{U}_i(\hat{\boldsymbol{\theta}})'] = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}_e\mathbf{X})$, and $\hat{\boldsymbol{\Sigma}}_e = \text{diag}[(y_i - \hat{\pi}_i)^2]$. The "bread" of the information sandwich estimate is the model estimated inverse information, or the estimated covariance matrix of the coefficient estimates. The "meat" of the sandwich is the *empirical* estimate of the observed information based on the empirical estimate of the error variance of the observations.

This robust estimate then can be used to construct confidence intervals and to compute Wald tests of significance. A robust score test of the model, or of model components, can also be constructed based on the model estimated under the tested hypothesis. The model score test addresses the null hypothesis $H_0: \boldsymbol{\beta} = \mathbf{0}$. The empirical estimate of the covariance matrix of the score vector is then obtained as

$$\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}_0) = \sum_i \mathbf{U}_i(\hat{\boldsymbol{\theta}}_0)\mathbf{U}_i(\hat{\boldsymbol{\theta}}_0)' = \mathbf{X}'\hat{\boldsymbol{\Sigma}}_{\varepsilon 0}\mathbf{X}, \quad (7.68)$$

where

$$\mathbf{U}_i(\hat{\boldsymbol{\theta}}_0) = [U_i(\hat{\boldsymbol{\theta}}_0)_\alpha \parallel \mathbf{U}_i(\hat{\boldsymbol{\theta}}_0)_\beta^T]^T \quad (7.69)$$

is the score vector with the parameters estimated ($\hat{\alpha}_0$) or evaluated (β) under the tested hypothesis as in Section 7.3.2. From (7.8) and (7.52) it follows that

$$U_i(\hat{\boldsymbol{\theta}}_0)_\alpha = y_i - \hat{\pi}, \quad (7.70)$$

and from (7.9) and (7.56), then

$$U_i(\hat{\boldsymbol{\theta}}_0)_{\beta_j} = x_{ij}(y_i - \hat{\pi}) \quad (7.71)$$

for the j th coefficient. Then the total score vector $\mathbf{U}(\hat{\boldsymbol{\theta}}_0)$ evaluated under the tested hypothesis is as presented in (7.57). Note that while the total score $U(\hat{\boldsymbol{\theta}})_\alpha = 0$, the individual terms in (7.69) are not all zero. The robust model score test then is

$$X^2 = \mathbf{U}(\hat{\boldsymbol{\theta}}_0)' \hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}_0)^{-1} \mathbf{U}(\hat{\boldsymbol{\theta}}_0), \quad (7.72)$$

that is asymptotically distributed as chi-square on p df.

Example 7.8 DCCT Nephropathy Data (continued)

The analysis of the DCCT prevalence of microalbuminuria in Table 7.6 contains the estimated covariance matrix of the coefficient estimates, $\hat{\Sigma}(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$. These and the estimated coefficients can be output to a data set by PROC LOGISTIC using the *outest* option. The program will also create a data set with the model estimated probabilities, the $\{\hat{\pi}_i\}$. Using a routine written in IML, additional computations may then be performed. The matrix of score vector outer products is $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}) = \sum_i [\mathbf{U}_i(\hat{\boldsymbol{\theta}}) \mathbf{U}_i(\hat{\boldsymbol{\theta}})'] = \mathbf{X}' \hat{\Sigma}_\epsilon \mathbf{X} =$

24.555087	9.1094437	237.04689	235.26848	2906.3223	9.5713568
9.1094437	9.1094437	88.132987	90.047355	1058.265	3.8137591
237.04689	88.132987	2351.74	2254.7264	27920.217	98.190859
235.26848	90.047355	2254.7264	2513.004	27998.372	92.930592
2906.3223	1058.265	27920.217	27998.372	347070.83	1105.9189
9.5713568	3.8137591	98.190859	92.930592	1105.9189	9.5713568

This yields the robust information sandwich estimate $\hat{\Sigma}_R(\hat{\boldsymbol{\theta}}) =$

10.664905	0.1689527	-0.267092	-0.05978	-0.063752	-0.374126
0.1689527	0.1965141	-0.02177	0.0012571	-0.000466	0.0515362
-0.267092	-0.02177	0.0243438	0.0021586	0.0002403	-0.017867
-0.05978	0.0012571	0.0021586	0.0041104	0.0000171	-0.001376
-0.063752	-0.000466	0.0002403	0.0000171	0.0005139	0.0039347
-0.374126	0.0515362	-0.017867	-0.001376	0.0039347	0.2162162

This can be used to compute large sample 95% confidence intervals for the model estimates and a Wald test of significance for the individual coefficients. The following are the vectors of estimated standard errors for the respective coefficients in the model,

the resulting upper and lower 95% confidence limits for the odds ratios, the Wald test chi-square value, and the p -value

Effect	S.E.($\hat{\theta}$)	95% Confidence Limits		Wald X^2	$p \leq$
		Lower	Upper		
Intercept	3.2657	-14.6810	-1.8800	6.429	0.0113
Intensive group	0.4433	-2.4520	-0.7143	12.750	0.0004
HbA _{1c}	0.1560	0.2617	0.8733	13.230	0.0003
Years duration	0.0641	-0.1160	0.1353	0.022	0.8809
Systolic B.P.	0.0227	-0.0211	0.0677	1.055	0.3043
Female	0.4650	-1.8019	0.0208	3.668	0.0555

For some of the coefficients, particularly those for intensive treatment group and HbA_{1c}, the robust standard errors are slightly larger than those estimated from the model presented in Table 7.6, so that the confidence limits for the odds ratios are slightly wider and the Wald tests slightly smaller.

The above supplemental computations are provided by the program *LogisticRobust.sas*. However, the robust information sandwich variance estimate, and the resulting robust confidence intervals and tests, are also provided by PROC GENMOD using the following statements

```
PROC Genmod data=renal descending; class subject;
  model micro24 = int hbael yearsdm sbp female
    / covb link=logit dist=bin;
  repeated subject=subject / type=unstr covb;
```

This is described in more detail in Section 8.4.2 in the context of Poisson regression.

This robust variance estimate can also be used to compute an overall model Wald test of $H_0: \beta = 0$. The value of the test is $X^2 = 21.34$ on 5 df with $p \leq 0.0007$. However, it is preferable that a score test be used that employs a robust estimate of the variance under the tested hypothesis. In this case, the score vector $U(\hat{\theta}_0)$ evaluated under H_0 from (7.57) is presented in Example 7.6, and the robust estimate of the covariance (information) matrix from (7.68) is $\hat{J}(\hat{\theta}_0) =$

31.744186	10.934694	309.35822	295.69036	3732.7345	12.325311
10.934694	10.934694	105.21842	103.05156	1263.9045	4.4315846
309.35822	105.21842	3092.0635	2845.4577	36218.088	126.01731
295.69036	103.05156	2845.4577	3058.0105	34860.286	113.10115
3732.7345	1263.9045	36218.088	34860.286	442423.16	1407.7262
12.325311	4.4315846	126.01731	113.10115	1407.7262	12.325311

The resulting robust score test is $X^2 = 33.64$ on 5 df with $p \leq 0.0001$.

Overall, for this model, inferences based on the robust information sandwich are nearly identical to those based on the simple model-based estimates of the expected

information. There is every indication that the model-based inferences are indeed appropriate. However, Chapter 8 presents an example where this is not the case.

7.3.6 Power and Sample Size

The power of a Wald test for the overall model or for an individual covariate in a logistic regression model analysis is a function of the noncentrality parameter of the test statistic. For a test of the coefficients in a model with a specific coefficient vector \mathbf{X} , Whittemore (1981) shows that the noncentrality parameter of the Wald test is a function of the moment generating function (*mgf*) of the joint multivariate distribution of \mathbf{X} . Because the expression for this *mgf* is, in general, unknown, it is not practical to describe the power function for a set of covariates, especially including one or more quantitative covariates.

7.3.6.1 Univariate Effects With a single binary covariate in the model, as shown in Chapter 6, the Wald test is approximately equivalent to the chi-square test for two proportions. Thus, the sample size required to detect a specific risk difference or odds ratio can be assessed from the power function of the test for two proportions as described in Chapter 3.

With a single quantitative covariate in the model, Hsieh et al. (1998) describe a simple method for the evaluation of sample size or power. One way to compare the quantitative variable values for positive and negative responses is to examine the difference between the means of these two response categories. When the mean difference is zero, it is readily shown that the estimated coefficient β for the covariate in a logistic regression model is also zero. This suggests that the large sample test for the difference between the means of two groups could be used to assess the sample size and power of a quantitative covariate in the logistic model.

Let π denote the probability of a positive response (or a case, $Y = 1$) for the binary outcome in the population and $1 - \pi$ the probability of a negative response (or control, $Y = 0$). Assume that the covariate (X) is distributed with means μ_1 and μ_0 among the positive and negative responses, with common variance σ^2 . Then the sample mean difference ($d = \bar{x}_1 - \bar{x}_2$) is asymptotically distributed as normal with mean $\Delta = \mu_1 - \mu_0$ and variance

$$V(d) = \sigma^2 = \frac{\pi^2}{N} \left(\frac{1}{\pi} + \frac{1}{1 - \pi} \right) = \frac{\pi^2}{N} \left(\frac{1}{\pi(1 - \pi)} \right), \quad (7.73)$$

that can be factored as $\sigma^2 = \phi^2/N$. Substituting into (3.12) yields the basic relationship

$$\frac{\sqrt{N} |\Delta|}{\phi} = \frac{Z_{1-\alpha} + Z_{1-\beta}}{\sqrt{\pi(1 - \pi)}}. \quad (7.74)$$

Hsieh et al. (1998) then equate the logistic regression model effect size under the alternative H_1 : $\beta = \beta_1$ with that of the test for means ($\beta_1 = \Delta/\phi$) to yield the

following expressions

$$Z_{1-\beta} = \sqrt{N\pi(1-\pi)}\beta_1 - Z_{1-\alpha} \quad (7.75)$$

$$N = \left[\frac{Z_{1-\alpha} + Z_{1-\beta}}{\beta_1 \sqrt{\pi(1-\pi)}} \right]^2.$$

Example 7.9 *Study of Post-traumatic Stress Disorder*

Hsieh et al. (2000) describe the sample size evaluation for a study of the association of heart rate responses to a sensory stimulus and the diagnosis of posttraumatic stress disorder. In the Veterans Administration patient population under evaluation, about 25% receive a positive diagnosis, or $\pi = 0.25$. To provide 90% power to detect a standardized difference in the mean heart rate response, or a log odds ratio, $\beta_1 = 0.3$, using a two-sided test at the 0.05 level, the sample size required from the above equation is 771.

7.3.6.2 Multiple Categorical Effects In the general case of a model with multiple categorical variables, the power of the Wald test of any one effect, or a contrast among effects, is readily obtained from the noncentral distribution of the chi-square test (cf. Rochon, 1989). Assume that the covariate vector generates K distinct cells. For example, for S binary covariates, $K = 2^S$ and the model contains $p \leq K$ binary design effects. This yields the $K \times (p+1)$ design matrix \mathbf{X} , where the i th row is the covariate vector \mathbf{x}_i^T for the i th cell augmented by the constant (unity) for the intercept. In this case, the binomial logit model of Section 7.1.2 specifies that the probability of the response within the i th subpopulation or cell characterized by covariate values \mathbf{x}_i is the logistic function $\pi = e^{\alpha + \mathbf{x}_i^T \boldsymbol{\beta}} / (1 + e^{\alpha + \mathbf{x}_i^T \boldsymbol{\beta}})$ with coefficients $\boldsymbol{\beta} = (\beta_1 \cdots \beta_p)^T$ and intercept α .

Then assume that the i th cell has expected sample size $E(n_i) = N\zeta_i$, N being the total sample size and $\{\zeta_i\}$ being the sample fractions. Within the i th cell, the observed proportion with the index characteristic is p_i . Since $V(\log[p_i/(1-p_i)]) = [n_i\pi_i(1-\pi_i)]^{-1}$, and the cells are independent subsamples, then the covariance matrix is Ω/N where $\Omega = \text{diag}\{1/[\zeta_i\pi_i(1-\pi_i)]\}$ and $\Omega^{-1} = \text{diag}[\zeta_i\pi_i(1-\pi_i)]$. Note that Ω is obtained directly as a function of \mathbf{x}_i and $\boldsymbol{\theta}$, $i = 1, \dots, K$. Then the parameters can be estimated through weighted least squares such that

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{Y}, \quad (7.76)$$

and

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = \Sigma_{\hat{\boldsymbol{\theta}}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}/N. \quad (7.77)$$

From Section A.7.1, the Wald test for a linear hypothesis of the form $H_0: \mathbf{C}'\boldsymbol{\theta} = 0$, for a $r \times (p+1)$ matrix \mathbf{C}' of row rank r , is the form

$$X^2 = \hat{\boldsymbol{\theta}}' \mathbf{C} (\mathbf{C}' \Sigma_{\hat{\boldsymbol{\theta}}} \mathbf{C})^{-1} \mathbf{C}' \hat{\boldsymbol{\theta}} \quad (7.78)$$

on r df. Thus, the noncentrality parameter, or the expected value of the test statistic under the alternative hypothesis, is

$$\begin{aligned}\psi^2 &= \boldsymbol{\theta}' \mathbf{C} (\mathbf{C}' \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \mathbf{C})^{-1} \mathbf{C}' \boldsymbol{\theta} \\ &= N \boldsymbol{\theta}' \mathbf{C} (\mathbf{C}' (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{C})^{-1} \mathbf{C}' \boldsymbol{\theta} = N \tau^2,\end{aligned}\quad (7.79)$$

where τ^2 is the noncentrality factor. Since ψ^2 is a function of \mathbf{C} , \mathbf{X} , and $\boldsymbol{\theta}$, then sample size and power can readily be determined from the noncentral chi-square distribution as described in Section 3.4.

Finally, in the event that overdispersion is thought to apply, the above is readily modified. All that is required is to inflate the values of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ by the desired factor that would correspond to the value of the overdispersion scale parameter in an overdispersed quasi-likelihood model. Alternatively, if power is being computed based on observed results, then the model may be fit using the information sandwich estimate of the variances that may then be used to compute the noncentral factor and the power of the test.

Example 7.10 Stratified Analysis

Consider the determination of the sample size for a study comparing two treatments within three strata with the design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix},$$

where the corresponding coefficients $\boldsymbol{\theta} = (\alpha \beta_1 \beta_2 \beta_3)^T$ represent the effects for the intercept, stratum 2 versus 1, stratum 3 versus 1, and the active (1) versus the control (0) treatment, respectively. Thus, rows 1 and 2 represent the two groups within stratum 1, and so on. Assume that subjects are assigned at random to each treatment such that there are equal sample sizes for each treatment group within each stratum. Also, assume that the three strata comprise 15%, 50%, and 35% of the total population, respectively. Thus, the expected cell fractions are specified to be $\{\zeta_i\} = (0.075, 0.075, 0.25, 0.25, 0.175, 0.175)$. Under the alternative hypothesis we specify that the probabilities for each cell are $\{\pi_i\} = (0.75, 0.923, 0.70, 0.882, 0.65, 0.770)$ for which the associated parameter vector is $\boldsymbol{\theta} = (1.2412 - 0.3158 - 0.7680 0.9390)^T$.

With these probabilities, the actual odds ratios for the active versus control treatment within each stratum are 4.0, 3.2, and 1.8, respectively. However, the design matrix and parameter vector values $\boldsymbol{\theta}$ correspond to a model where the adjusted (common) odds ratio for active treatment versus control is $e^{\beta_3} = e^{0.9390} = 2.557$. Thus, sample size and power are evaluated under a fixed effects model with some heterogeneity, which provides a conservative assessment of sample size or power relative to the model with a common odds ratio for all strata.

Substituting the values of $\{\pi_i\}$ and $\{\zeta_i\}$ into (7.77) yields the covariance matrix of the estimates. For the test of $H_0: \beta_3 = 0$, with vector $\mathbf{C}' = (0 \ 0 \ 0 \ 1)$, the

resulting value of the noncentrality factor for the 1 df test is $\tau^2 = 0.03484$. For a 1 df Wald chi-square test at $\alpha = 0.05$ with a critical value of 3.84146, the SAS function CNONCT(3.84146, 1, 0.1) yields the noncentrality parameter value $\psi^2(0.05, 0.10, 1) = 10.5074$ required to provide power = 0.90. Thus, the total N required to provide 90% power for this test is $N = 10.5074/0.03484 = 302$ (rounded up from 301.6).

Conversely, for the same model, a sample size of only $N = 200$ provides a noncentrality parameter $\psi^2(0.05, \beta, 1) = N\tau^2 = (200 \times 0.0348427) = 6.969$. The SAS function PROBCHI(3.841, 1, 6.969) then yields $\beta = 0.248$ and power of 0.752.

For this and similar examples, one approach to specifying the parameter vector θ under the alternative is to first specify the $\{\pi_i\}$ that yield the desired treatment group odds ratios within each stratum (or subpopulation) under the alternative hypothesis. For the above example, the values of $\{\pi_i\}$ were obtained by first specifying the values in the control group within each stratum, (0.75, 0.70, and 0.65), and then specifying the odds ratio within each stratum (4.0, 3.2, and 1.8), which then yields the probabilities in the active treatment group within each stratum (0.923, 0.882, and 0.770). The parameter vector θ was then obtained by generating a set of cell frequencies summing to a large number, say 10,000, the frequencies being proportional to the corresponding cell fractions $\{\zeta_i\}$, and then fitting the logistic regression model to obtain the corresponding values of θ .

Alternatively, Rochon (1989) first uses the vector of probabilities $\{\pi_i\}$ and cell sample fractions $\{\zeta_i\}$ to define the covariance matrix Ω . The design matrix for the assumed model is then employed in (7.76) to obtain the expected vector of parameters θ . These correspond to the first-step estimates in an iteratively reweighted computation and thus are not precisely correct because the covariance matrix Ω is based on the specified alternative model in the $\{\pi_i\}$, not the model in θ ; that is, using the vector $\pi_\theta = 1/[1 - \exp(x'\theta)]$. However, the iteratively reweighted computation with Ω defined using the elements of π_θ will yield parameters θ and a sample size similar to those described above.

7.3.6.3 Models Adjusted for Multiple Factors In some cases, the objective of the analysis is to assess the effects of a single index risk factor (say X with coefficient β_X) adjusted for a number of other factors (say, Z with coefficient vector β_Z). In such cases, the variance of the index factor estimate will be increased and the power for the test of the index factor will be diminished in proportion to the extent of confounding or correlation with other factors. This suggests that a variance inflation factor could be used in conjunction with a univariate assessment. Let $\hat{\beta}_X$ and $V(\hat{\beta}_X)$ denote the coefficient estimate and its variance for the index risk factor in a univariate unadjusted model. Then let $\hat{\beta}_{X|Z}$ denote the coefficient estimate for X in a model adjusted for the other covariates Z , where $R_{X|Z}^2$ is the coefficient of determination for the regression of X on Z . For a logistic model with quantitative covariates, Whittemore (1981) showed that variance of the adjusted estimate was approximately

$$V(\hat{\beta}_{X|Z}) = V(\hat{\beta}_X)[1 - R_{X|Z}^2]^{-1}. \quad (7.80)$$

Since the sample size required to provide a given level of power is inversely proportional to $V(\hat{\beta}_X)$, this suggests that if the sample size computation had used the adjusted variance, the covariate-adjusted sample size, say N_A , would satisfy

$$N_A = \frac{N}{1 - R_{X|Z}^2}.$$

For the assessment of power, the expression for a univariate model would employ $N_A[1 - R_{X|Z}^2]$, where N_A is the target sample size. This adjustment has also been used in other settings, such as the Cox proportional hazards model by Hsieh and Lavori (2000).

Example 7.11 *Study of Post-traumatic Stress Disorder (continued)*

Hsieh et al. (2000) also describe the adjustment of the sample size for the effect of heart rate response (X) on the probability of a diagnosis of posttraumatic stress disorder when the analysis will be adjusted for three other factors (Z = blood pressure, EMG, and skin conductance). The regression of X on Z was expected to yield $R_{X|Z}^2 = 0.1$ and a variance inflation factor of 1.11. Thus, the previously calculated $N = 771$ would be inflated to yield $N_A = 857$.

Example 7.12 *Test of Proportions*

Example 3.2 showed that a sample size of 652 would provide 90% power to detect a difference in proportions of 0.4 versus 0.28 with two equally sized groups at the 0.05 level, two-sided. This also corresponds to the test of an odds ratio of 1.71 in a univariate logistic regression model. If the study were randomized, then the treatment group binary covariate would be expected to have zero correlation with other baseline factors, in which case no further adjustment would be required, even if a baseline covariate-adjusted analysis were planned. However, if the two groups were not randomized, as in a case-control study, a covariate-adjusted analysis would be appropriate, and the sample size for the group comparison should be adjusted accordingly using the above variance inflation factor.

7.3.6.4 SAS PROC POWER Finally, it should be noted that SAS PROC POWER also includes a *logistic* statement that will conduct sample size or power computations for a qualitative, ordinal, or quantitative covariate effect in a logistic model. The procedure also allows for adjustment for other covariates. However, all of the effects in the model are assumed to be statistically independent (i.e., uncorrelated). This is useful for a designed experiment, such as a factorial, where sample size or power is assessed for one factor adjusting for the other.

7.4 INTERACTIONS

Thus far, all of the models we have considered contain main effects only. In many instances, however, models that include interactions between covariates provide a better description of the risk relationships. In the simplest sense, an interaction implies that the effect of one covariate depends on the value of another covariate, and

vice versa. Such effects can either be coded directly or specified as an interaction effect in the model. When coded directly, a new variable must be defined in the data step, such as $x12=x1*x2$, where $x1*x2$ is the product of the values for X_1 and X_2 . The variable $x12$ is then used in the *model* statement, such as *model* $y=x1\ x2\ x12$. Alternatively, the interaction effect can be specified in a *model* statement such as *model* $y=x1\ x2\ x1*x2$. Interactions involving categorical variables are also readily fit using a *class* statement.

First, consider the simple linear regression model with an identity link, for example, where Y is a quantitative variable and $E(Y|\mathbf{X})$ is the conditional expectation for Y given the value of the covariate vector \mathbf{X} . The resulting model can be parameterized as

$$E(Y|\mathbf{X}) = \alpha + X_1\beta_1 + X_2\beta_2 + X_1X_2\beta_{12}, \quad (7.81)$$

where $X_1X_2\beta_{12}$ is the interaction term with coefficient β_{12} . When $\beta_{12} = 0$, the simple main-effects-only model applies. However, when $\beta_{12} \neq 0$, we have, in effect, two alternative representations of the same model:

$$\begin{aligned} E(Y|\mathbf{X}) &= \alpha + X_1\tilde{\beta}_{1|X_2} + X_2\beta_2, & \tilde{\beta}_{1|X_2} &= \beta_1 + X_2\beta_{12} \\ &= \alpha + X_1\beta_1 + X_2\tilde{\beta}_{2|X_1}, & \tilde{\beta}_{2|X_1} &= \beta_2 + X_1\beta_{12}. \end{aligned} \quad (7.82)$$

Thus, the coefficient (slope) of the effect for one covariate depends on the value of the other covariate, and vice versa, and the two covariates interact.

Note that in this case, the main effects β_1 and β_2 do not describe the complete effect of each covariate in the model. Rather, the complete effect of X_1 is represented by the values of β_1 and β_{12} , and the complete effect of X_2 by the values of β_2 and β_{12} . Thus, the tests of the main effects alone do not provide tests of the complete effects of X_1 and X_2 . Rather, a 2 *df* test of $H_0: \beta_1 = \beta_{12} = 0$ is required to test the overall effect of X_1 , and likewise a 2 *df* test of $H_0: \beta_2 = \beta_{12} = 0$ is required to test the overall effect of X_2 . These tests are not computed automatically by any standard program, but some programs, including LOGISTIC, provide a *test* option to conduct multiple-degree-of-freedom Wald tests of hypotheses.

In simple regression, the coefficients are interpreted in terms of the change in the expected value of the mean given a unit change in the covariates. In logistic regression, these effects are interpreted in terms of the change in the log odds or the log odds ratio.

7.4.1 Qualitative–Qualitative Covariate Interaction

The simplest case is where the two covariates X_1 and X_2 are both binary. As in Example 7.1, let $O_{\mathbf{x}}$ designate the odds of the positive outcome for a subject with covariate values \mathbf{x} , that includes the X_1X_2 interaction term. Then

$$O_{\mathbf{x}} = \exp(\alpha + \mathbf{X}'\boldsymbol{\beta}) = \exp(\alpha + X_1\beta_1 + X_2\beta_2 + X_1X_2\beta_{12}). \quad (7.83)$$

There are only four categories of subjects in the population, each with odds given by

Category	x_1	x_2	x_1x_2	O_x
1	0	0	0	e^α
2	0	1	0	$e^{\alpha+\beta_2}$
3	1	0	0	$e^{\alpha+\beta_1}$
4	1	1	1	$e^{\alpha+\beta_1+\beta_2+\beta_{12}}$

(7.84)

Note that if there is no interaction term, $\beta_{12} = 0$, then in the last cell the effects are additive in the exponent, or multiplicative.

In this case, it is instructive to consider the odds ratio of each category versus the first, designated as $OR_{(x_1, x_2):(0,0)}$, as presented in the cells of the following table

		$OR_{(x_1, x_2):(0,0)}$		
		x_2		
		0	1	$OR_{(x_1, 1):(x_1, 0)}$
x_1	0	1.0	e^{β_2}	e^{β_2}
	1	e^{β_1}	$e^{\beta_1+\beta_2+\beta_{12}}$	$e^{\beta_2+\beta_{12}}$
$OR_{(1, x_2):(0, x_2)}$		e^{β_1}	$e^{\beta_1+\beta_{12}}$	

(7.85)

The odds ratios associated with each covariate, conditional on the value of the other, are obtained by the ratios within rows and within columns. These are given along the row and column margins of the table. The variance of each log odds ratio is then obtained from the variance of the corresponding linear combination of the estimated coefficients. For example, $V[\log \widehat{OR}_{(1,1):(0,1)}] = V(\widehat{\beta}_1 + \widehat{\beta}_{12}) = V(\widehat{\beta}_1) + V(\widehat{\beta}_{12}) + 2\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_{12})$. From these, confidence limits and Wald tests may then be computed.

Example 7.13 Coronary Heart Disease in the Framingham Study (continued)

Using the data from Example 7.1, a logistic model with an interaction yields the following model coefficients

Effect	No Interaction			Interaction		
	$\widehat{\beta}$	\widehat{OR}	$p \leq$	$\widehat{\beta}$	\widehat{OR}	$p \leq$
$X_1: \text{HiChol}$	1.0301	2.801	0.0001	1.1751	3.238	0.0012
$X_2: \text{HiSBP}$	0.9168	2.501	0.0001	1.1598	3.189	0.0114
$X_1 X_2$	—	—	—	-0.3155	0.729	0.5459

For comparison, the no-interaction model from Table 7.4 is also presented. The interaction effect is not statistically significant, indicating that the additive exponential model applies. Nevertheless, it is instructive to describe the interaction model.

To assess the overall significance of each effect, a 2 *df* test could be obtained using the following statements as part of the SAS PROC LOGISTIC analysis:

```
test hichol=interact=0;
test hisbp=interact=0;
```

The resulting test of the effect of high cholesterol yields a 2 *df* chi-square test value of $X^2 = 15.7577$ with $p \leq 0.0004$, and the test for the effect of high blood pressure yields $X^2 = 17.7191$ with $p \leq 0.0001$.

In such instances, the 1 *df* tests computed by the program and presented above should not be used as tests of the overall effect of each covariate. Rather, they are tests of the simple effect of each covariate when added to the interaction. In some cases, these simple 1 *df* effects may not be significant, whereas the test of interaction and the 2 *df* overall test are significant. This would indicate that the simple effect of the covariate does not add significantly to the model in the presence of the interaction effect.

From this interaction model we can construct a table of odds ratios comparable to (7.85) as follows:

		$OR_{(x_1, x_2):(0,0)}$		$OR_{(x_1, 1):(x_1, 0)}$	(7.86)
		0	1		
$x_1: HiChol$	0	1.0	3.189	3.189	
	1	3.238	7.527	2.325	
$OR_{(1, x_2):(0, x_2)}$		3.238	2.360		

Thus, the odds ratio associated with high versus low cholesterol is 3.238 for those with low blood pressure and 2.360 for those with high blood pressure. Likewise, the odds ratio associated with high versus low blood pressure is 3.189 for those with low cholesterol and 2.325 for those with high cholesterol. Note that the odds ratios associated with the main effects in the interaction model (3.189 and 3.238) are those for each covariate, where the value of the other covariate is zero.

For comparison, the following table provides the odds ratios derived from the no-interaction model:

		$OR_{(x_1, x_2):(0,0)}$		$OR_{(x_1, 1):(x_1, 0)}$	(7.87)
		0	1		
$x_1: HiChol$	0	1.0	2.501	2.501	
	1	2.801	7.007	2.501	
$OR_{(1, x_2):(0, x_2)}$		2.801	2.801		

In this case the odds ratios for one covariate are constant for the values of the other covariate. Also, the effects of the covariates are multiplicative or exponentially additive, where $OR_{(1,1):(0,0)} = e^{x_1\beta_1 + x_2\beta_2} = (2.501 \times 2.801) = 7.007$. Thus, the

test of the interaction effect can be viewed as a test of the difference between the odds ratios in the (1,1) cell, or between the 7.007 under the no-interaction model versus the 7.527 under the interaction model.

Example 7.14 Ulcer Clinical Trial: Stratum \times Group Interaction

Table 4.4 presents a SAS program for the analysis of the Ulcer Clinical Trial data of Example 4.1 stratified by ulcer type. The following statements when added to this program would fit a logistic regression model that includes a treatment \times stratum interaction:

```
data three; set two;
z2=0; if k=2 then z2=1; iz2=i*z2;
z3=0; if k=3 then z3=1; iz3=i*z3
proc logistic descending;
model j = i z2 z3 iz2 iz3 / rl;
weight f;
test iz2=iz3=0;
```

Alternatively, the interaction model could be fit using a *class* statement with interaction effects such as

```
proc logistic descending;
class i z2 z3 / param=ref descending;
model j = i z2 z3 i*z2 i*z3 / rl;
weight f;
```

In the following we employ the former direct coding.

The covariates z_2 and z_3 represent the effects of strata 2 and 3, respectively, versus the reference stratum 1, and iz_2 and iz_3 represent the treatment group \times covariate interaction. The fitted model is of the form

$$\begin{aligned}\log(\hat{O}_x) &= \hat{\alpha} + i\hat{\beta}_1 + z_2\hat{\beta}_2 + z_3\hat{\beta}_3 + (iz_2)\hat{\beta}_{12} + (iz_3)\hat{\beta}_{13} \\ &= -0.3001 - i(0.1854) + z_2(0.077) - z_3(0.2595) \\ &\quad + (iz_2)(1.5072) + (iz_3)(1.1869),\end{aligned}\tag{7.88}$$

where β_2 and β_3 are equivalent to the γ_2 and γ_3 of Section 7.1.4.

In this model, all covariates are binary, so that $e^{\hat{\alpha}} = 0.7407$ is the estimated odds of healing within the category of the population with all covariates equal to 0, or for the control group of stratum 1. Then $e^{\hat{\alpha}+\hat{\beta}_1} = 0.6154$ is the odds of healing in the drug-treated group within the first stratum. Their ratio is the estimate of the odds ratio for the drug versus placebo within this stratum, $\widehat{OR}_1 = 0.831$. Likewise, the odds within each group within each stratum, and the odds ratio within each stratum, are obtained from the fitted model as

Stratum	z_2	z_3	<i>Placebo</i>	<i>Drug</i>	<i>Drug:Placebo</i>	(7.89)
			<i>Odds</i> ($i = 0$)	<i>Odds</i> ($i = 1$)	<i>Odds Ratio</i>	
1	0	0	0.741	0.615	0.831	
2	1	0	0.800	3.000	3.750	
3	0	1	0.571	1.556	2.722	

These model-based estimates of the treatment group odds ratios within each stratum are identical to those within each stratum as shown in Table 4.1. Converting these odds back to the corresponding proportions yields the observed proportions within each cell of the data structure. Such models are called *saturated* because they provide an exact fit to the observed data and no other parameters can be added to the model without the addition of more covariates to the data structure.

Clearly, if the two interaction terms equal zero, then the model reduces to the stratified-adjusted model of Section 7.1.4, that provides an estimate of the common odds ratio for treatment group within each stratum. Thus, a test of no interaction on 2 *df* provides a test of homogeneity comparable to those described in Sections 4.6.1 and 4.6.2. A Wald test of $H_0: \beta_{12} = \beta_{13} = 0$ is provided by the *test* option in PROC LOGISTIC. The resulting test yields $X^2 = 4.5803$ with $p \leq 0.1013$, which is equivalent to the value of the Cochran test of homogeneity presented in Example 4.10.

Alternatively, the likelihood ratio test can be computed as the difference in the model X^2 tests for the interaction versus no-interaction model to yield $X^2 = 11.226 - 6.587 = 4.639$ with $p \leq 0.0983$ on 2 *df*. This test can also be obtained from PROC GENMOD using a model with class effects for group and stratum and with the *type3* option.

7.4.2 Interactions with a Quantitative Covariate

Similar considerations apply to the interpretation of a model with covariate effects between a qualitative and a quantitative covariate, or between two quantitative covariates. These cases are illustrated in the following examples.

Example 7.15 DCCT Nephropathy: Treatment by Duration Interaction

A model fit to the DCCT nephropathy data of Example 7.4 with interactions between treatment and duration of diabetes, and also between the HbA_{1c} and level of systolic

blood pressure, yields the following:

Effect	$\hat{\beta}$	\widehat{OR}	$p \leq$	
X_1 : <i>Intensive Treatment</i>	-2.4410	0.087	0.0697	
X_2 : <i>HbA_{1c}</i>	3.7478	42.428	0.0182	
X_3 : <i>Years Duration of DM</i>	-0.0103	0.990	0.8997	(7.90)
X_4 : <i>Systolic Blood Pressure</i>	0.2752	1.317	0.0297	
X_5 : <i>Female</i>	-0.9505	0.387	0.0386	
X_1X_3 : <i>Intensive Rx \times Duration</i>	0.0772	1.080	0.5636	
X_2X_4 : <i>HbA_{1c} \times SBP</i>	-0.0272	0.973	0.0431	

Note that the model does not include an interaction with gender, so that the female effect can be interpreted as a true main effect. The other covariates each involve an interaction, and thus the combined effect of each covariate is contained in part in the main effect and in part in the interaction term.

The treatment group \times duration interaction is not statistically significant. Nevertheless, it is instructive to describe the interpretation of such an interaction between a qualitative and a quantitative covariate. Here, one must consider a specific value (or values) of the quantitative covariate. For example, consider the odds of microalbuminuria in each treatment group for subjects with nine and 10 years duration of diabetes:

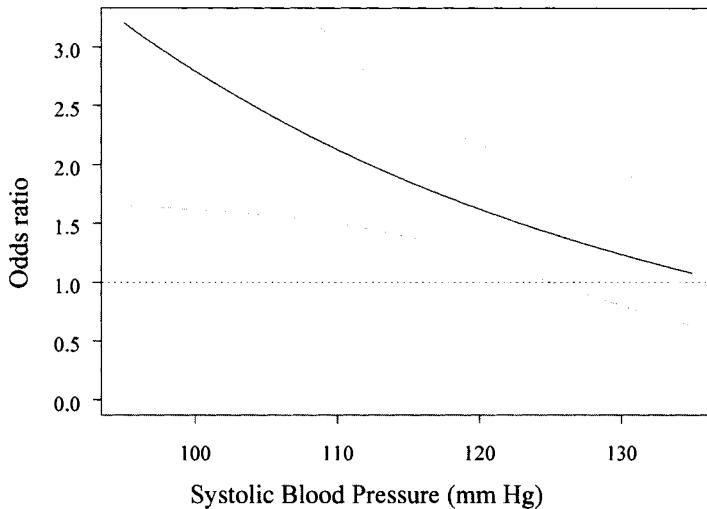
Group	Duration	\widehat{O}_x	
1 (Intensive)	9	$Ce^{\hat{\beta}_1 + 9(\hat{\beta}_3 + \hat{\beta}_{13})}$	$= C(0.15899)$
0 (Conv.)	9	$Ce^{9\hat{\beta}_3}$	$= C(0.91147)$
1 (Intensive)	10	$Ce^{\hat{\beta}_1 + 10(\hat{\beta}_3 + \hat{\beta}_{13})}$	$= C(0.16999)$
0 (Conv.)	10	$Ce^{10\hat{\beta}_3}$	$= C(0.90213)$

where the constant $C = \exp(x_2\hat{\beta}_2 + x_4\hat{\beta}_4 + x_5\hat{\beta}_5 + x_2x_4\hat{\beta}_{24})$ for any fixed values of the other covariates. This yields odds ratios for intensive versus conventional therapy for those with nine years duration of $\widehat{OR}_{I:C|D=9} = (0.15899/0.91147) = 0.1744 = \exp(\hat{\beta}_1 + 9\hat{\beta}_{13})$; and for those with 10 years duration of $\widehat{OR}_{I:C|D=10} = 0.1884 = \exp(\hat{\beta}_1 + 10\hat{\beta}_{13})$. Likewise, within the intensive therapy group, the odds ratio associated with one year longer duration, such as 10 versus nine years, is $\widehat{OR}_{D+1|I} = (0.16999/0.15899) = 1.06919 = \exp(\hat{\beta}_3 + \hat{\beta}_{13})$; and in the conventional group, $\widehat{OR}_{D+1|C} = \exp(\hat{\beta}_3) = 0.990$, as shown above.

Similar calculations for other values of duration would allow estimation of the odds ratio for intensive versus conventional therapy over a range of values of duration. Such computations are described in the next example.

Note that neither interaction involves the covariates in the other interaction. Thus, the effects of covariates X_1 to X_4 each involve only one other covariate. If a covariate is involved in multiple interactions, then the effect of that covariate depends on the values of all other covariates included in any of these interactions. This complicates

Fig. 7.1 Odds ratio per unit higher value of HbA_{1c} (%) as a function of the level of systolic blood pressure in mmHg.



the interpretation of the results, but using simple generalizations of the above, it is possible to describe the effects of each covariate as a function of the values of the other covariates in higher dimensions.

Example 7.16 DCCT Nephropathy: HbA_{1c} × Blood Pressure Interaction

In the above model, the Wald test for the interaction between HbA_{1c} (say, H) and systolic blood pressure (say, S) is statistically significant at the 0.05 level. This indicates that the effect of one variable depends explicitly on the value of the other. Since the biological hypothesis is that blood sugar levels, as represented by the HbA_{1c}, are the underlying causal mechanism, it is more relevant to examine the effect of HbA_{1c} over a range of values of blood pressure than vice versa. For a given value of systolic blood pressure (x_4), the odds ratio associated with a one unit (%) increase in HbA_{1c} is provided by

$$\widehat{OR}_{H+1|S} = \exp \left[\widehat{\beta}_2 + x_4 \widehat{\beta}_{24} \right]. \quad (7.92)$$

This function can be plotted over a range of values of blood pressure (X_4) as shown in Figure 7.1. For each value of blood pressure (x_4) the 95% asymmetric confidence limits for this odds ratio can be computed from those based on the linear combination of the estimates, where

$$\widehat{V}(\widehat{\beta}_2 + x_4 \widehat{\beta}_{24}) = \widehat{V}(\widehat{\beta}_2) + x_4^2 \widehat{V}(\widehat{\beta}_{24}) + 2x_4 \widehat{Cov}(\widehat{\beta}_2, \widehat{\beta}_{24}). \quad (7.93)$$

The variances and covariances (not shown) can be obtained from the *covb* option in PROC LOGISTIC.

As the level of baseline blood pressure increases, the odds ratio associated with a unit increase in the HbA_{1c} decreases. For a moderately low blood pressure, such as 110 mmHg, the odds ratio for a unit increase in HbA_{1c} is 2.13; for an average blood pressure of about 116, the HbA_{1c} odds ratio is 1.81, close to that in the no-interaction model in Table 7.6; and for a moderately high blood pressure, such as 120 mmHg, the HbA_{1c} odds ratio is 1.62.

7.5 MEASURES OF THE STRENGTH OF ASSOCIATION

There are three central objectives in data analysis. The first is to describe the *nature* of the association among the independent variables or covariates and the dependent variable. In a logistic regression model this is expressed by the coefficient estimates $\hat{\beta}$ or the corresponding odds ratios and their confidence limits. The second is to assess the *statistical significance* of the association by conducting a test of a relevant hypothesis concerning the values of the coefficients, such as the test of the model null hypothesis $H_0: \beta = 0$. The third is to describe the *strength* of the association using some measure of the fraction of variation in the dependent variable that is explained by the covariates through the estimated model relationship.

7.5.1 Squared Error Loss

Section A.8.2 describes the use of the squared error loss function to quantify the fraction of squared error loss explained, ρ_{ε^2} . In a logistic regression model, where y_i is an indicator variable for the response (D) versus not (\bar{D}), this is expressed as a measure of the explained variation in risk, $\rho_{\varepsilon^2}^2 = \rho_{risk}^2$ (Korn and Simon, 1991). Unconditionally, with no covariate effects, or under the null model, $E(y_i) = P(D) = \bar{\pi}$ and $V(y) = \bar{\pi}(1 - \bar{\pi})$. Alternatively, the logistic regression model specifies that the probability of the response conditional on covariates is $E(y_i | \mathbf{x}_i) = P(D | \mathbf{x}_i) = \pi_i(\mathbf{x}_i) = \pi_i$, where $\pi_i = [1 + e^{-\mathbf{x}_i \boldsymbol{\theta}}]^{-1}$. Then

$$V[E(y|\mathbf{x})] = E(\pi_i - \bar{\pi})^2 \quad \hat{=} \quad \frac{1}{N} \sum_i (\hat{\pi}_i - \bar{\hat{\pi}})^2, \quad (7.94)$$

where $\bar{\hat{\pi}} = m_1/N$. Therefore, the estimated explained risk with squared error loss from (A.190) is

$$\hat{\rho}_{\varepsilon^2}^2 = \hat{\rho}_{risk}^2 = \frac{\hat{V}[\hat{E}(y|\mathbf{x})]}{\hat{V}(y)} = \frac{\frac{1}{N} \sum_i (\hat{\pi}_i - \bar{\hat{\pi}})^2}{\bar{\hat{\pi}}(1 - \bar{\hat{\pi}})}. \quad (7.95)$$

When the logistic regression model fits perfectly, then $\hat{\pi}_i = 1.0$ for all subjects where $y_i = 1$ (or for whom the response occurred), and $\hat{\pi}_i = 0$ for all subjects where

$y_i = 0$. In this case it is readily shown that $\hat{\rho}_{risk}^2 = 1.0$. Conversely, when the model explains none of the variation in risk, then $\hat{\pi}_i = \bar{\pi}$ for all subjects, irrespective of the covariate values and $\hat{\rho}_{risk}^2 = 0$ (see Problem 7.14).

As described in Section A.8.2, the fraction of squared error loss explained by the model is also estimated by the fraction of residual variation. In logistic regression it is readily shown that $R_{\varepsilon^2, resid}^2 = \hat{\rho}_{risk}^2$ above.

7.5.2 Entropy Loss

Although squared error loss pervades statistical practice, other loss functions have special meaning in some situations. The entropy loss in logistic regression and the analysis of categorical data is one such instance, as described by Efron (1978).

Consider a discrete random variable Y with probability mass function $P(y)$, $0 \leq P(y) \leq 1$ for $y \in S$, the sample space of Y . The certainty or uncertainty with which one can predict an individual observation depends on the higher moments of $P(y)$ or the degree of dispersion in the data. One measure of this predictive uncertainty or dispersion is the entropy of the distribution of Y (Shannon, 1948) defined as

$$H(Y) = - \sum_{y \in S} P(y) \log[P(y)] = -E(\log[P(y)]). \quad (7.96)$$

For example, assume that the sample space consists of the set of K integers $y \in S = (1, \dots, K)$ with equal probability $P(y) = 1/K$ for $y \in S$. Then the entropy for this distribution is

$$H(Y) = - \sum_{i=1}^K \left(\frac{1}{K} \right) \log \left(\frac{1}{K} \right) = - \log \left(\frac{1}{K} \right) = \log(K). \quad (7.97)$$

In fact, this is the maximum possible entropy $H(Y)$ for a discrete random variable with K categories. Thus, for this distribution there is the least possible certainty with which one can predict any single observation drawn at random compared to any other member of the family of K -element discrete distributions.

Conversely, if the sample space consists of a single integer, say $y = b$, where $P(y = b) = 1$ and $P(y) = 0$ for $y \neq b$, then the entropy for this point mass distribution is $H(Y) = 1 \log(1) = 0$. This is the minimum entropy possible for any distribution, discrete or continuous.

Now consider a Bernoulli random variable such as the indicator variable for the response in a logistic regression model. Here Y is an indicator variable with sample space $S = (0, 1)$ and where $E(Y) = P(Y = 1) = \pi$. Then the entropy for this distribution is

$$\begin{aligned} H(Y) &= - \sum_{y=0}^1 [y\pi \log \pi + (1-y)(1-\pi) \log(1-\pi)] \\ &= - [\pi \log(\pi) + (1-\pi) \log(1-\pi)] \\ &= -E(\log[p(y)]). \end{aligned} \quad (7.98)$$

This suggests that an *entropy loss function* can be defined for any one random observation as

$$\mathcal{L}_E(y, \hat{\pi}) = -[y \log(\hat{\pi}) + (1 - y) \log(1 - \hat{\pi})] \quad (7.99)$$

for any prediction function for the Bernoulli probability of that observation based on covariates such that $\hat{\pi} = P(D | \mathbf{x})$ (see Problem 7.14.3). Then for a population of observations, each with a predicted or estimated probability $\hat{\pi}_i = P(D | \mathbf{x}_i)$, the expected loss is the entropy for the corresponding Bernoulli probabilities

$$D_E(\mathbf{x}) = E[\mathcal{L}_E(y_i, \hat{\pi}_i)] = -[\pi_i \log(\pi_i) + (1 - \pi_i) \log(1 - \pi_i)] = H(y|\pi_i). \quad (7.100)$$

Now under the null model with no covariates, all observations are identically distributed as Bernoulli with probability $\bar{\pi}$, and the expected entropy is

$$D_{E0} = E[\mathcal{L}_E(y, \bar{\pi})] = -[\bar{\pi} \log(\bar{\pi}) + (1 - \bar{\pi}) \log(1 - \bar{\pi})] = H(y|\bar{\pi}). \quad (7.101)$$

Thus, the expected fraction of entropy loss explained, ρ_E^2 , is obtained by substituting these quantities into (A.180).

This fraction of entropy explained can then be estimated as

$$\hat{\rho}_E^2 = \frac{\hat{D}_{E0} - \hat{D}_E(\mathbf{x})}{\hat{D}_{E0}}, \quad (7.102)$$

using the sample estimate of the entropy loss based on the estimated logistic function with covariates

$$\hat{D}_E(\mathbf{x}) = -\frac{1}{N} \left[\sum_i \hat{\pi}_i \log(\hat{\pi}_i) + (1 - \hat{\pi}_i) \log(1 - \hat{\pi}_i) \right], \quad (7.103)$$

and that unconditionally without covariates

$$\hat{D}_{E0} = -\left[\hat{\pi} \log(\hat{\pi}) + (1 - \hat{\pi}) \log(1 - \hat{\pi})\right]. \quad (7.104)$$

Therefore,

$$\hat{\rho}_E^2 = 1 - \frac{\frac{1}{N} \sum_i [\hat{\pi}_i \log(\hat{\pi}_i) + (1 - \hat{\pi}_i) \log(1 - \hat{\pi}_i)]}{\hat{\pi} \log(\hat{\pi}) + (1 - \hat{\pi}) \log(1 - \hat{\pi})}. \quad (7.105)$$

Alternatively, the fraction of explained entropy loss can be estimated from the fraction of residual variation explained in terms of the entropy loss function, which from (A.199) is

$$\begin{aligned} R_{E,resid}^2 &= \frac{\sum_i \mathcal{L}_E(y_i, \hat{\pi}) - \sum_i \mathcal{L}_E(y_i, \hat{\pi}_i)}{\sum_i \mathcal{L}_E(y_i, \hat{\pi})} \\ &= 1 - \frac{\sum_i [y_i \log(\hat{\pi}_i) + (1 - y_i) \log(1 - \hat{\pi}_i)]}{\sum_i [y_i \log(\hat{\pi}) + (1 - y_i) \log(1 - \hat{\pi})]}. \end{aligned} \quad (7.106)$$

In Problem 7.14 it is then shown that these estimates are identical, $\hat{\rho}_E^2 = R_{E,resid}^2$, based on the fact that the score equation for the intercept $U(\boldsymbol{\theta})_\alpha$ implies that $\sum_i y_i = \sum_i \hat{\pi}_i = N\bar{\pi}$, and that for each coefficient $U(\boldsymbol{\theta})_{\beta_j}$ implies that $\sum_i y_i x_{ij} = \sum_i \hat{\pi}_i x_{ij}$, $j = 1, \dots, p$.

The latter expression for $R_{E,resid}^2$ is especially meaningful in logistic regression since the two expressions in (7.106) refer to logistic regression log likelihood functions, such that

$$R_{E,resid}^2 = 1 - \frac{\log L(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})}{\log L(\hat{\boldsymbol{\alpha}})} = \frac{X_{LR}^2}{-2 \log L(\hat{\boldsymbol{\alpha}})} = R_\ell^2 \quad (7.107)$$

where X_{LR}^2 is the model likelihood ratio chi-square test value on p df.

The latter quantity R_ℓ^2 has a natural interpretation as the fraction of $-2 \log L$ explained by the covariates in a logistic regression model. However, the interpretation as a fraction of entropy loss explained by the model is applicable only to regression models with a Bernoulli response, such as the logistic regression model. Nevertheless, the explained log likelihood has become widely used in other regression models, such as Poisson regression, since the null model $-2 \log L(\hat{\boldsymbol{\alpha}})$ for any model is analogous to the unconditional $SS(Y)$. Further, as shown in Section A.8.3, any model based on a likelihood with distribution $f(y|x)$ implies a loss function equal to $-\log[f(y|x)]$ that then yields a fraction of explained loss equal to R_ℓ^2 .

Another measure of explained variation described in Section A.8.4 is Madalla's R_{LR}^2 , which is also a function of the model likelihood ratio statistic. However, this measure differs from the entropy R_E^2 . From (7.107), stated more generally in terms of a null model with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then

$$R_E^2 = 1 - \frac{\log L(\hat{\boldsymbol{\theta}})}{\log L(\boldsymbol{\theta}_0)} \neq R_{LR}^2 = 1 - \exp[-X_{LR}^2/N]. \quad (7.108)$$

Also, Nagelkerke (1991) shows that the likelihood ratio explained variation $R_{LR}^2 < 1$ for models based on discrete probability distributions, such as the binomial or Bernoulli in logistic regression, where $\max(R_{LR}^2) = 1 - L(\boldsymbol{\theta}_0)^{2/N}$. Thus, he suggested that the R_{LR}^2 be rescaled so as to achieve a value of 1.0 when R_{LR}^2 is at its maximum. Thus, the maximum rescaled R_{LR}^2 is computed as

$$\max \text{ rescaled } R_{LR}^2 = \frac{R_{LR}^2}{\max(R_{LR}^2)} = \frac{1 - \exp\left[-\frac{X_{LR}^2}{N}\right]}{1 - L(\boldsymbol{\theta}_0)^{2/N}} = \frac{1 - \left[\frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})}\right]^{2/N}}{1 - L(\boldsymbol{\theta}_0)^{2/N}}. \quad (7.109)$$

These R_{LR}^2 measures are generated using the *rsquare* option in PROC LOGISTIC. However, the program does not compute the other R^2 measures based on squared error loss, $\hat{\rho}_{risk}^2 = R_{\epsilon^2,resid}^2$, or on the entropy, $\hat{\rho}_E^2 = R_{E,resid}^2$.

These expressions can also be employed to describe the partial R^2 associated with each individual covariate effect in a model. For a given covariate effect, the covariate

X^2 is substituted for the model X_{LR}^2 , either from the Wald test, or preferably, from the likelihood ratio test for the covariate. As is generally true in multiple regression models, the covariate effect R^2 measures will not sum to the overall model R^2 . Rather, each covariate partial R^2 is a measure of the change in the model R^2 when that effect is dropped from the model, approximately so when the Wald test is employed.

Example 7.17 DCCT Nephropathy Data (continued)

To evaluate $\hat{p}_{risk}^2 = R_{\varepsilon^2, resid}^2$ for the model in Table 7.6, the estimated probabilities for all observations must be output into a data set and that the numerator sum of squares be calculated, such as using PROC UNIVARIATE. By definition, from the score estimating equation for the intercept, $\sum_i \hat{\pi}_i = 42 = m_1$ which is the number of responses, so that $\bar{\pi} = 42/172 = 0.244186$. From the computed estimated probabilities of all observations, $\sum_i (\hat{\pi}_i - \bar{\pi})^2 = 6.662041$. Therefore,

$$R_{\varepsilon^2, resid}^2 = \frac{6.662041}{(172)(0.244186)(1 - 0.244186)} = 0.20987.$$

For these data, $-2 \log L(\hat{\alpha}) = -2 \log L(\theta_0) = 191.215$ and the model likelihood ratio chi-square test is $X_{LR}^2 = 35.879$. Thus, the explained entropy loss or negative log likelihood is $R_E^2 = 35.879/191.215 = 0.18764$. Similarly, Madalla's likelihood ratio-based measure is $R_{LR}^2 = 1 - \exp[-35.879/172] = 0.18828$. This is the value of *RSquare* computed using the *rsquare* option as shown in Table 7.6. The program also presents the *Max-rescaled RSquare* = 0.2806. The latter is inflated relative to the other measures and has no obvious interpretation in terms of a loss function. The $R_{\varepsilon^2, resid}^2$ and R_E^2 are clearly preferable.

7.6 CONDITIONAL LOGISTIC REGRESSION MODEL FOR MATCHED SETS

7.6.1 Conditional Logistic Model

Section 6.6.2 introduces the conditional logit model for the analysis of the 2×2 table from a prospective sample of matched pairs of exposed and nonexposed individuals. Section 6.6.5 also shows that this model can be applied to the analysis of a retrospective sample of matched pairs of cases and controls. Those developments apply to the special case of 1:1 or pair matching with a single binary covariate (exposure or case group).

Breslow (1982) describes the general conditional logistic regression model to assess the effects of covariates on the odds of the response in studies that involve matched sets of multiple exposed and nonexposed individuals in a prospective study, or multiple cases and controls in a retrospective study. This model allows for a vector of covariates for each subject and also for each pair. (See the discussion of pair and member stratification in Section 5.7.1.) The model is described herein in terms of a prospective study but may also be applied to a retrospective study.

Consider the case of N matched sets of size n_i for the i th set, $i = 1, \dots, N$. The j th member of the i th set has covariate vector $\mathbf{x}_{ij} = (x_{ij1} \ \dots \ x_{ijp})^T$ for $i = 1, \dots, N$; $j = 1, \dots, n_i$. For each member one or more of these covariates describe the exposure status of that member, either qualitatively, such as exposed versus not, or quantitatively, such as the dose and the duration of exposure. The other covariates may then represent pair characteristics that are shared by all members of the matched set, or characteristics that are unique to each member of each set (see the discussion of pair and member stratification in Section 5.7). Then for the ij th member, let y_{ij} be an indicator variable representing whether the member experienced or developed the positive response of interest ($D : y_{ij} = 1$) or did not ($\bar{D} : y_{ij} = 0$). We then adopt a logistic model which specifies that the probability of the response π_{ij} for the ij th member is a logistic function of the covariates \mathbf{x}_{ij} . The model allows each matched set to have a unique risk of the response represented by a set-specific intercept α_i , and then assumes that the covariates have a common effect on the odds of the response over all matched sets represented by the coefficient vector $\beta = (\beta_1 \ \dots \ \beta_p)^T$. Thus, the model is of the form

$$P_i(D|\mathbf{x}_{ij}) = \pi_{ij} = \pi_i(\mathbf{x}_{ij}) = \frac{e^{\alpha_i + \mathbf{x}'_{ij}\beta}}{1 + e^{\alpha_i + \mathbf{x}'_{ij}\beta}}, \quad (7.110)$$

and

$$\log \left[\frac{\pi_{ij}}{1 - \pi_{ij}} \right] = \alpha_i + \mathbf{x}'_{ij}\beta. \quad (7.111)$$

Now consider two members from any matched set (the i th), one with covariate vector \mathbf{x} , the other \mathbf{x}_0 . Then the odds ratio for these two members is

$$\frac{\pi_i(\mathbf{x}) / (1 - \pi_i(\mathbf{x}))}{\pi_i(\mathbf{x}_0) / (1 - \pi_i(\mathbf{x}_0))} = \frac{e^{\alpha_i + \mathbf{x}'\beta}}{e^{\alpha_i + \mathbf{x}'_0\beta}} = \exp [(\mathbf{x} - \mathbf{x}_0)' \beta]. \quad (7.112)$$

Thus, the odds ratio for two members of any other set with covariate values differing by the amount $\mathbf{x} - \mathbf{x}_0$ is assumed to be constant for all matched sets regardless of the value of α_i .

Within the i th set the n_i members are conditionally independent. Thus, the unconditional likelihood in terms of the parameters $\alpha = (\alpha_1 \ \dots \ \alpha_N)^T$ and β is of the form

$$L(\alpha, \beta) = \prod_{i=1}^N \prod_{j=1}^{n_i} \pi_i(\mathbf{x}_{ij})^{y_{ij}} [1 - \pi_i(\mathbf{x}_{ij})]^{1-y_{ij}}, \quad (7.113)$$

which yields the log likelihood

$$\begin{aligned} \ell(\alpha, \beta) &= \sum_{i=1}^N \sum_{j=1}^{n_i} y_{ij} \log [\pi_i(\mathbf{x}_{ij})] + (1 - y_{ij}) \log [1 - \pi_i(\mathbf{x}_{ij})] \quad (7.114) \\ &= \sum_i \sum_j \left[y_{ij} \log \left(\frac{e^{\alpha_i + \mathbf{x}'_{ij}\beta}}{1 + e^{\alpha_i + \mathbf{x}'_{ij}\beta}} \right) - (1 - y_{ij}) \log \left(1 + e^{\alpha_i + \mathbf{x}'_{ij}\beta} \right) \right] \\ &= \sum_i m_{1i} \alpha_i + \sum_i \sum_j \left[y_{ij} \mathbf{x}'_{ij}\beta - \log \left(1 + e^{\alpha_i + \mathbf{x}'_{ij}\beta} \right) \right] \end{aligned}$$

where m_{1i} denotes the number of members with a positive response ($y_{ij} = 1$) and m_{2i} those without ($y_{ij} = 0$) in the i th matched set. As was the case with a single covariate to represent exposed versus not exposed as in Section 6.6, to obtain an estimate of the covariate vector β requires that we also solve for the N nuisance parameters $\{\alpha_i\}$. However, this can be avoided by conditioning on the number of positive responses m_{1i} within the i th matched set that is the sufficient statistic for the nuisance parameter α_i for that set.

To do so, it is convenient to order the members in the i th set such that the first m_{1i} subjects are those with a response. Now, the conditional probability of interest for the i th matched set is the probability of m_{1i} responses with covariates $\mathbf{x}_1, \dots, \mathbf{x}_{m_{1i}}$ and m_{2i} nonresponses with covariates $\mathbf{x}_{(m_{1i}+1)}, \dots, \mathbf{x}_{n_i}$ given m_{1i} responses out of n_i members, or

$$P \left[\left\{ (m_{1i}; \mathbf{x}_1, \dots, \mathbf{x}_{m_{1i}}) \cap (m_{2i}; \mathbf{x}_{(m_{1i}+1)}, \dots, \mathbf{x}_{n_i}) \right\} \mid m_{1i}, n_i \right]. \quad (7.115)$$

Thus, the conditional likelihood for the i th matched set is

$$L_{i|m_{1i}, n_i} = \frac{\prod_{j=1}^{m_{1i}} \pi_i(\mathbf{x}_{ij}) \prod_{j=m_{1i}+1}^{n_i} [1 - \pi_i(\mathbf{x}_{ij})]}{\sum_{\ell=1}^{\binom{n_i}{m_{1i}}} \prod_{j(\ell)=1}^{m_{1i}} \pi_i(\mathbf{x}_{ij(\ell)}) \prod_{j(\ell)=m_{1i}+1}^{n_i} [1 - \pi_i(\mathbf{x}_{ij(\ell)})]}. \quad (7.116)$$

The denominator is the sum over all possible combinations of m_{1i} of the n_i subjects with the response and m_{2i} subjects without. Within the ℓ th combination $j(\ell)$ then refers to the original index of the member in the j th position.

Substituting the logistic function for $\pi_i(\mathbf{x}_{ij})$, it is readily shown that

$$L_{i|m_{1i}, n_i} = \frac{\prod_{j=1}^{m_{1i}} e^{\mathbf{x}'_{ij} \beta}}{\sum_{\ell=1}^{\binom{n_i}{m_{1i}}} \prod_{j(\ell)=1}^{m_{1i}} e^{\mathbf{x}'_{ij(\ell)} \beta}}. \quad (7.117)$$

Therefore, for the sample of N matched sets, the conditional likelihood is

$$L_{(c)}(\beta) = \prod_i L_{i|m_{1i}, n_i} \quad (7.118)$$

and the log likelihood is

$$\ell_{(c)}(\beta) = \sum_i \ell_{i|m_{1i}, n_i}(\beta), \quad (7.119)$$

where

$$\begin{aligned}\ell_{i|m_{1i}, n_i}(\boldsymbol{\beta}) &= \sum_{j=1}^{m_{1i}} \mathbf{x}'_{ij} \boldsymbol{\beta} - \log \left[\sum_{\ell=1}^{\binom{n_i}{m_{1i}}} \prod_{j(\ell)=1}^{m_{1i}} e^{\mathbf{x}'_{ij(\ell)} \boldsymbol{\beta}} \right] \\ &= \mathbf{s}'_i \boldsymbol{\beta} - \log \left[\sum_{\ell=1}^{\binom{n_i}{m_{1i}}} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}} \right],\end{aligned}\quad (7.120)$$

where $\mathbf{s}_i = \sum_{j=1}^{m_{1i}} \mathbf{x}_{ij}$ for the i th set and $\mathbf{s}_{i(\ell)} = \sum_{j(\ell)=1}^{m_{1i}} \mathbf{x}_{ij(\ell)}$ for the ℓ th combination of m_{1i} subjects within that set. The resulting score vector is $\mathbf{U}(\boldsymbol{\beta}) = [U(\boldsymbol{\beta})_{\beta_1} \ \dots \ U(\boldsymbol{\beta})_{\beta_p}]^T$, where the equation for the k th element of the coefficient vector is

$$U(\boldsymbol{\beta})_{\beta_k} = \sum_{i=1}^N \frac{\partial \ell_{i|m_{1i}, n_i}(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^N \left[s_{ik} - \frac{\sum_{\ell} s_{i(\ell)k} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}}{\sum_{\ell} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}} \right], \quad (7.121)$$

where s_{ik} and $s_{i(\ell)k}$ denote the k th element of \mathbf{s}_i and $\mathbf{s}_{i(\ell)}$, respectively, and \sum_{ℓ} denotes $\sum_{\ell=1}^{\binom{n_i}{m_{1i}}}$. The expected information matrix $\mathbf{I}(\boldsymbol{\beta})$ then has diagonal elements ($1 \leq k \leq p$)

$$\mathbf{I}(\boldsymbol{\beta})_{\beta_k} = \sum_{i=1}^N \left[\left(\frac{\sum_{\ell} s_{i(\ell)k}^2 e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}}{\sum_{\ell} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}} \right) - \left(\frac{\sum_{\ell} s_{i(\ell)k} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}}{\sum_{\ell} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}} \right)^2 \right], \quad (7.122)$$

and off-diagonal elements ($1 \leq k < u \leq p$)

$$\begin{aligned}\mathbf{I}(\boldsymbol{\beta})_{\beta_k, \beta_u} &= \sum_{i=1}^N \left(\frac{\sum_{\ell} s_{i(\ell)k} s_{i(\ell)u} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}}{\sum_{\ell} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}} \right) \\ &\quad - \sum_{i=1}^N \left(\frac{\sum_{\ell} s_{i(\ell)k} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}}{\sum_{\ell} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}} \right) \left(\frac{\sum_{\ell} s_{i(\ell)u} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}}{\sum_{\ell} e^{\mathbf{s}'_{i(\ell)} \boldsymbol{\beta}}} \right).\end{aligned}\quad (7.123)$$

Note that since the coefficient estimates $\hat{\boldsymbol{\beta}}$ are derived from a conditional likelihood, with no intercept, then this model describes the relative odds as a function of the covariates (\mathbf{x}_{ij}), not the absolute probabilities or risks $\{\pi_i(\mathbf{x}_{ij})\}$. To model the absolute risks, we would have to use the full unconditional likelihood (7.114), in which we must also estimate the nuisance parameters $(\alpha_1, \dots, \alpha_N)$.

7.6.2 Matched Retrospective Study

Now consider a matched retrospective case-control study with N matched sets consisting of n_i members of whom m_{1i} are cases ($y_{ij} = 1$) and m_{2i} are controls ($y_{ij} = 0$). Each member of each set then has an associated covariate vector \mathbf{x}_{ij} ,

some elements of which may represent prior exposure variables. As described by Breslow et al. (1978), we then wish to model the posterior (retrospective) probabilities for cases and controls:

$$\phi_{i1}(\mathbf{x}) = P_i(\mathbf{x}|y=1) \quad \text{and} \quad \phi_{i2}(\mathbf{x}) = P_i(\mathbf{x}|y=0) \quad (7.124)$$

described in Section 6.6.5. We again apply a conditioning argument to the i th matched set to assess the conditional probability as in (7.115) of m_{1i} covariate vectors $\mathbf{x}_{i1}, \dots, \mathbf{x}_{im_{1i}}$ among m_{1i} cases and m_{2i} covariate vectors $\mathbf{x}_{i(m_{1i}+1)}, \dots, \mathbf{x}_{in_i}$ among m_{2i} controls given m_{1i} cases among the matched set of n_i subjects. For the i th matched set this yields the conditional likelihood

$$\tilde{L}_{i|m_{1i},n_i} = \frac{\prod_{j=1}^{m_{1i}} \phi_i(\mathbf{x}_{ij}) \prod_{j=m_{1i}+1}^{n_i} [1 - \phi_i(\mathbf{x}_{ij})]}{\sum_{\ell=1}^{\binom{n_i}{m_{1i}}} \prod_{j(\ell)=1}^{m_{1i}} \phi_i(\mathbf{x}_{ij(\ell)}) \prod_{j(\ell)=m_{1i}+1}^{n_i} [1 - \phi_i(\mathbf{x}_{ij(\ell)})]}. \quad (7.125)$$

However, as was the case for a single 2×2 table in Section 6.6.5, each of the retrospective probabilities ϕ_{ij} then can be expressed as a function of the prospective probabilities π_{ij} . Substituting into (7.125) it is readily shown that

$$\tilde{L}_{i|m_{1i},n_i} = L_{i|m_{1i},n_i}, \quad (7.126)$$

as presented in (7.116) and (7.117). Therefore, the conditional logistic regression model can be applied to a matched retrospective study to describe the effects of covariates on the prospective odds of the disease, or of being a case.

7.6.3 Fitting the General Conditional Logistic Regression Model

As described in earlier releases of SAS, the procedure PHREG can be used to fit the conditional logistic model since the likelihood is identical to that for the Cox proportional hazards model with ties. This requires construction of a dummy variable to represent "time" in the model. However, later releases of SAS provide direct computation of the model with PROC LOGISTIC using a stratum specification, as illustrated in the following examples.

7.6.4 Allowing for Clinic Effects in a Randomized Trial

A common application of the conditional logistic model is to assess the difference among treatment groups adjusting for possible heterogeneity among the participating clinical centers in a multicenter study. Often a clinical trial may involve tens or hundreds of sites. In this case, it can be highly inefficient to include a clinic class effect as a fixed effect in an unconditional logistic model. As an alternative, a conditional logistic model can effectively assess the treatment group effect adjusted for clinic as a stratifying variable.

Example 7.18 DCCT Treatment Group Adjusted for Clinic

Example 7.4 describes the assessment of the effects of intensive versus conventional treatment on the risk of microalbuminuria at six years of follow-up in a subset of the DCCT. It might also be of interest to assess the magnitude of the treatment effect adjusted for clinical center. The prior example employed a subset of the subjects for convenience. For these analyses, the entire secondary intervention cohort is employed. Thus, the results in this section are not comparable to those presented previously. The complete data set (*clinren*) is available from the book website.

The DCCT was initiated in 21 clinical centers. Some years afterward, additional centers were added, but there were few subjects enrolled in these centers in time to be followed for six or more years. Thus, that set of clinics has been combined (clinic # 2242). Among the original 21 clinics, no subjects were observed to have microalbuminuria in clinic 18. Thus, clinic 18 was arbitrarily combined with clinic 19 (clinic # 1819).

In the case of a sparse clinic, or one with no positive outcomes, it is preferable that the clinic in question be retained by combining it with other like clinics. In the case of clinic 18, the zero positive values could have arisen as a function of the characteristics of the patients at that site, such as where all were absent risk factors for microalbuminuria. In this case, it could bias the results to exclude them.

To begin, consider the fixed effects model for the effect of treatment group adjusted for other factors. In this larger data set, the model, like that in Table 7.6, provides

Model Fit Statistics			
Criterion	-2 Log L	Intercept Only	Intercept and Covariates
	479.402	438.141	

with a model likelihood ratio test chi-square value of 41.2609 on 5 *df* with *p* < 0.0001. The parameter estimates are

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Odds Ratio
Intercept	1	-6.8740	1.6705	16.9321	<.0001	
INT	1	-0.6195	0.2469	6.2979	0.0121	0.538
HBAEL	1	0.4218	0.0795	28.1843	<.0001	1.525
YEARSMD	1	0.0978	0.0361	7.3380	0.0068	1.103
SBP	1	0.00798	0.0116	0.4772	0.4897	1.008
FEMALE	1	-0.5772	0.2717	4.5116	0.0337	0.561

We will focus principally on the difference between the treatment groups with coefficient estimate -0.6195 and a corresponding odds ratio of 0.538 for intensive versus conventional therapy; i.e., a 46.2% reduction in the odds. The 95% confidence limits on the parameter estimate provide confidence limits on the odds ratio of (0.332, 0.873).

To adjust for clinic as a class (fixed) effect, the following statements would be employed

```
proc logistic descending;
  class clinic;
  model micro24 = int hbael yearsdm sbp female clinic / rl;
```

This results in a model with $-2 \log(L) = 415.775$ and a model chi-square test value of 63.6267 on 25 *df*. Compared to the model above, the likelihood ratio test of the added clinic effects yields a chi-square of 22.37 on 20 *df* that is clearly not significant. The adjusted coefficient for treatment group (*int*) is -0.6066 with *S.E.* = 0.2562 and *p* = 0.0179. The odds ratio and 95% confidence limits are 0.545 (0.330, 0.901). Compared to the above model not adjusted for clinic, the point estimate and odds ratio are minimally affected and the confidence limits are slightly wider, owing to a slight increase in the coefficient *S.E.*.

An alternative approach to adjust for clinic would be to fit a conditional regression model stratified by clinic using the statements

```
proc logistic descending;
  strata clinic;
  model micro24 = int hbael yearsdm sbp female / rl;
```

See the program *clinrenal.sas*. The results are shown in Table 7.9.

First note that the $-2 \log L$ value without covariates is 389.091, that differs from that above using the unconditional logistic model. Therefore, the fit of a conditional versus unconditional model cannot be compared in this manner. The model likelihood ratio chi-square test value is 35.1115 on 5 *df*, *p* < 0.0001. The treatment group coefficient is -0.5778 with *S.E.* = 0.2500 and *p* = 0.0208. This yields an odds ratio and confidence limits of 0.561 (0.344, 0.916). The odds ratio is slightly closer to 1, but the *S.E.* is about the same as above, so that the resulting *p*-value is slightly larger.

Both the unconditional model with clinic added as a fixed effect and the conditional model stratifying by clinic assume that the risk of microalbuminuria varies among clinics but that there is a common odds ratio comparing intensive versus conventional therapy within clinic. With a moderate number of clinics, the choice of model is probably irrelevant. However, with a large number of clinics (or strata), the conditional model is preferred. Bartlett (1955) showed that a model with a large number of nuisance parameters estimated by maximum likelihood provides a biased estimate of the other parameters in the model, such as an estimate of the treatment effect when adjusted for a large number of clinics. Various authors have proposed corrections for this bias. See Liang and Zeger (1995) and Efron (1998). Hu and Lachin (2003) present a simulation to describe the bias and the adequacy of the various corrections. However, in a conditional logistic model, little bias is introduced because the model does not attempt to provide an explicit estimate of the parameter associated with each clinic.

Table 7.9 Conditional logistic regression analysis of DCCT microalbuminuria data stratified by clinical center.

The LOGISTIC Procedure
Conditional Analysis

Strata Summary					
Response	MICRO24		Number of Strata	Frequency	
Pattern	1	0			
1	4	14	1		18
2	3	16	1		19
3	4	15	1		19
4	4	16	1		20
5	5	15	1		20
6	6	14	1		20
7	7	13	1		20
8	3	18	1		21
9	2	21	1		23
10	4	19	1		23
11	1	23	1		24
12	4	20	1		24
13	6	20	1		26
14	8	19	1		27
15	2	26	1		28
16	6	23	1		29
17	5	25	1		30
18	3	28	1		31
19	5	28	1		33
20	2	34	1		36
21	4	43	1		47

Model Fit Statistics					
		Without		With	
Criterion		Covariates		Covariates	
-2 Log L		389.091		353.979	

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard	Wald	
				Chi-Square	Pr > ChiSq
INT	1	-0.5778	0.2500	5.3415	0.0208
HBAEL	1	0.3902	0.0804	23.5241	<.0001
YEARSDM	1	0.0961	0.0373	6.6288	0.0100
SBP	1	0.00637	0.0125	0.2591	0.6108
FEMALE	1	-0.6369	0.2836	5.0419	0.0247

Later, in Section 7.8, mixed effects models are described in which clinic effect, and possibly treatment group within clinic, enter the model as random effects.

Example 7.19 *Low Birth Weight*

Hosmer and Lemeshow (2004) present data from a study of factors associated with the risk of giving birth to a low-weight infant. Table 7.10 presents a subset of the data (reproduced with permission). These data are also presented in the SAS (1997) manual and online documentation that describes the use of the program PROC PHREG to fit the conditional logistic model for matched sets.

Sets of pregnant women were matched according to the year of age, the set sizes ranging from three to 18 women. The risk factors considered (in addition to age) were the mother's weight (in pounds) at her last menstrual period (*lwt*), an indicator variable for whether or not the mother smoked (*smoke*, 1 = yes, 0 = no), whether she had hypertension (*ht*, 1 = yes, 0 = no), and whether there was evidence of uterine irritation (*ui*, 1 = yes, 0 = no).

The data in Table 7.10 are organized by matched sets indicated by the variable *matchset*, each set comprising all women with the same age (16 years for *matchset* 1, 17 for *matchset* 2, etc.). This includes the number of each case and control (*casen* and *controln*) within each matched set, the covariate values, and the indicator variable for whether or not the mother gave birth to a low-birth-weight infant (*case*, 1=yes, 0=no). This data set is provided in the author's program *HLbirthweight.sas*. The complete data set, not arranged by matched set, is available in the above references.

Table 7.11 then presents the results of the conditional logistic regression analysis of these data, extraneous material deleted. There are 17 matched sets with a total of 174 mothers, of whom 54 (31%) gave birth to a low-birth-weight infant. These were age-matched to 120 controls who gave birth to a normal-birth-weight infant. The conditional logistic model yields a model likelihood ratio test of 17.961 on 4 *df* with $p \leq 0.0013$. Although this is a matched case-control study, the parameter estimates equal the estimated prospective log odds ratios for the risk factors. Women who smoked and who had hypertension had significantly ($p \leq 0.05$) increased odds of having a low-birth-weight infant. Women with uterine irritation had an estimated 2.4-fold increased odds, but this was not statistically significant. The odds of a low-birth-weight infant decreased by 1.5% per pound higher maternal body weight. When expressed in terms of a 10-lb increase in body weight, there is an estimated odds ratio of 0.86, or a 14% risk reduction. Alternatively, when expressed in terms of a 10-lb decrease, there is an estimated $OR = 1.162$ or a 16.2% increase in risk.

7.6.5 Robust Inference

SAS PROC GENMOD does not provide an option for a stratified analysis and thus cannot be used to provide robust inferences for a conditional logistic model. However, robust inferences can be obtained using the information sandwich for the Cox proportional hazards model described by Lin and Wei (1989) and Lin (1994), that are provided by the SAS PROC PHREG, as described in Section 9.6.4. This would require the construction of a dummy time variable coded as *time* = 2 - *case*,

Table 7.10 SAS program for conditional logistic regression analysis of low-birth-weight data from Hosmer and Lemeshow (1989), reproduced with permission.

```

data one;
  input matchset casen controln lwt smoke ht ui case;
  cards;
  1   1   0   130   0   0   0   0   1
  1   0   1   110   0   0   0   0   0
  1   0   2   112   0   0   0   0   0
  1   0   3   135   1   0   0   0   0
  1   0   4   135   1   0   0   0   0
  1   0   5   170   0   0   0   0   0
  1   0   6   95    0   0   0   0   0
  2   1   0   130   1   0   1   1   1
  2   2   0   110   1   0   0   0   1
  2   3   0   120   1   0   0   0   1
  2   4   0   120   0   0   0   0   1
  2   5   0   142   0   1   0   0   1
  2   0   1   103   0   0   0   0   0
  2   0   2   122   1   0   0   0   0
  2   0   3   113   0   0   0   0   0
  2   0   4   113   0   0   0   0   0
  2   0   5   119   0   0   0   0   0
  2   0   6   119   0   0   0   0   0
  2   0   7   120   1   0   0   0   0
  .
  .
  .
  .
  17   1   0   105   1   0   0   0   1
  17   0   1   121   0   0   0   0   0
  17   0   2   132   0   0   0   0   0
  17   0   3   134   1   0   0   0   0
  17   0   4   170   0   0   0   0   0
  17   0   5   186   0   0   0   0   0
  ;
  proc sort; by matchset;
  proc logistic descending;
    model case = lwt smoke ht ui;
    strata matchset;
  run;

```

Table 7.11 Conditional logistic regression analysis of low-birth-weight data.

The LOGISTIC Procedure					
Strata Summary					
Response Pattern	Case		Number of Strata	Frequency	
	1	0			
1	2	1	1	3	
2	1	4	1	5	
3	1	5	1	6	
4	1	6	3	21	
5	4	4	1	8	
6	2	7	1	9	
7	2	8	1	10	
8	5	7	2	24	
9	2	11	1	13	
10	5	8	2	26	
11	6	9	1	15	
12	3	13	1	16	
13	8	10	1	18	

Model Fit Statistics					
Criterion	Without Covariates		With Covariates		
-2 Log L	159.069		141.108		

Testing Global Null Hypothesis: BETA=0					
Test	Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio	17.9613	4	0.0013		
Score	17.3152	4	0.0017		
Wald	15.5578	4	0.0037		

Analysis of Maximum Likelihood Estimates						
Variable	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Odds Ratio
LWT	1	-0.0150	0.0071	4.5002	0.0339	0.985
SMOKE	1	0.8080	0.3680	4.8222	0.0281	2.244
HT	1	1.7514	0.7393	5.6120	0.0178	5.763
UI	1	0.8834	0.4803	3.3827	0.0659	2.419

that equals 1 if a case, 2 a control. Then the following statements would fit the conditional PH model with the same results as presented in Table 7.11:

```
proc phreg;
  model time*case(0) = lwt smoke ht ui / ties = discrete;
  strata matchset;
```

Then additional specifications and computations would provide the empirical covariance estimates and tests of significance described in Section 9.6.4.

7.6.6 Explained Variation

For the unconditional logistic regression model, based on the expression for the log likelihood in (7.114), it follows that the fraction of explained log likelihood equals the fraction of explained entropy, or $R_E^2 = R_\ell^2$. However, in the conditional logistic model, based on the expression for the log likelihood in (7.119), the fraction of explained log likelihood, R_ℓ^2 , although still valid, does not have a direct interpretation as a fraction of explained entropy loss. Also, since the conditional logistic model likelihood equals that of the Cox proportional hazards model, measures of explained variation for the PH model may also be applied to the conditional logistic model. In a simulation study, Schemper (1992) showed that Madalla's likelihood ratio measure R_{LR}^2 (see Section 9.4.8) provides a simple approximation to more precise measures of explained variation for the PH model. In the conditional logistic model stratified by matched set, this measure is computed as

$$R_{LR}^2 = 1 - \exp \left[\frac{-X_{LR}^2}{\sum_{i=1}^N n_i} \right], \quad (7.127)$$

where the sample size in this expression is actually the sum total number of members.

In the above example, the fraction of explained log likelihood is

$$R_\ell^2 = 17.961/159.069 = 0.129; \quad (7.128)$$

whereas Madalla's $R_{LR}^2 = 1 - \exp [-17.961/174] = 0.098$.

7.6.7 Power and Sample Size

Schoenfeld (1983) derived an expression for the power of a score test of a covariate in the Cox PH model, as described later in Section 9.5.2. Lachin (2008) likewise described the assessment of power and sample size for the score test of a quantitative or binary covariate effect in a conditional logistic regression model with possibly variable numbers of cases and controls. He first derived the expression for the score test $Z = U(\beta_0)/\sqrt{I(\beta_0)}$ for a single covariate in the model for any number of cases and controls within each matched set. Denoting $s_i = \sum_{j=1}^{m_{1i}} x_{ij}$ for the i th set and $s_{i(\ell)} = \sum_{j(\ell)=1}^{m_{1i}} x_{ij(\ell)}$ for the ℓ th of $\binom{n_i}{m_{1i}}$ combinations of m_{1i} subjects within

that set, then

$$U(\beta_0) = \sum_{i=1}^N \left[s_i - \frac{\sum_{\ell} s_{i(\ell)}}{\binom{n_i}{m_{1i}}} \right] = \sum_{i=1}^N [s_i - E(s_i|H_0)], \quad (7.129)$$

and

$$I(\beta_0) = \sum_{i=1}^N \left[\left(\frac{\sum_{\ell} s_{i(\ell)}^2}{\binom{n_i}{m_{1i}}} \right) - \left(\frac{\sum_{\ell} s_{i(\ell)}}{\binom{n_i}{m_{1i}}} \right)^2 \right] = \sum_{i=1}^N I_{0i}. \quad (7.130)$$

Tang (2009) shows that these expressions can be simplified to yield

$$U(\beta_0) = \sum_{i=1}^N \left(\frac{m_{1i}m_{2i}}{n_i} \right) (\bar{x}_{i1} - \bar{x}_{i2}), \quad (7.131)$$

$$\begin{aligned} I(\beta_0) &= \sum_{i=1}^N I_{0i} = \sum_{i=1}^N \left(\frac{m_{1i}m_{2i}}{n_i} \right) \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1} \\ &= \sum_{i=1}^N \left(\frac{m_{1i}m_{2i}}{n_i} \right) \hat{\sigma}_i^2, \end{aligned}$$

where $\bar{x}_{i1} = \sum_{j=1}^{m_{1i}} x_{ij}/m_{1i}$ and $\bar{x}_{i2} = \sum_{j=1}^{m_{2i}} x_{ij}/m_{2i}$ are the covariate means among the cases and controls, respectively, and $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ is the mean for the i th set.

Under the alternative hypothesis that the covariate coefficient is nonzero, $H_1: \beta = \beta_1 \neq 0$, then the score equation is distributed as

$$U(\beta_0)|H_1 \sim N[\beta_1 I(\beta_0), I(\beta_1)], \quad (7.132)$$

and

$$Z|H_1 \sim \mathcal{N} \left[\beta_1 \sqrt{I(\beta_0)}, \left(\frac{I(\beta_1)}{I(\beta_0)} \right) \right] \quad (7.133)$$

(see Section 9.5.2). For a specific true value $\beta = \beta_1$ under the alternative hypothesis, the basic equation for the power of the test yields

$$|\beta_1| \sqrt{I(\beta_0)} = Z_{1-\alpha} + Z_{1-\beta} \sqrt{I(\beta_1)/I(\beta_0)} \quad (7.134)$$

$$Z_{1-\beta} = \frac{|\beta_1| \sqrt{I(\beta_0)} - Z_{1-\alpha}}{\sqrt{I(\beta_1)/I(\beta_0)}} \cong |\beta_1| \sqrt{I(\beta_0)} - Z_{1-\alpha}$$

where $Z_{1-\alpha}$ is the critical value for the test, using $Z_{1-\alpha/2}$ for a two-sided test, and the power of the test is provided by $\Phi(Z_{1-\beta})$, $\Phi(z)$ being the standard normal cdf at z . Equations for specific instances are then derived by solving for the expression for $I(\beta_0)$. Herein the resulting expressions are presented for a study with fixed numbers of positive outcomes (or cases, $m_{1i} = m_1$) and negative outcomes (or controls, $m_{2i} = m_2$) for all sets (i) (see Lachin (2008) and Tang (2009) for more general expressions).

7.6.7.1 Quantitative Covariate Consider the case of a single quantitative covariate X with common variance σ^2 among N matched sets with m_1 cases and m_2 controls so that within each set $V(s_i) = V(s_{i(\ell)}) = m_1\sigma^2$. Using Tang's simplification in (7.131), it follows that

$$I_{0i} = \frac{m_1 m_2 \sigma^2}{n} \quad (7.135)$$

for all strata (i), and

$$I(\beta_0) = N \frac{m_1 m_2 \sigma^2}{n}. \quad (7.136)$$

This result also applies to the heteroscedastic case for suitably large N such that $\sum_i \sigma_i^2 \rightarrow N\sigma^2$, where $E(\sigma_i^2) = \sigma^2$.

For a model with coefficient β_1 , the power of the score test is provided by

$$Z_{1-\beta} = |\beta_1| \sigma \left[\frac{Nm_1 m_2}{n} \right]^{1/2} - Z_{1-\alpha}. \quad (7.137)$$

The β_1 is the log odds ratio per unit change in X of interest to detect. Thus, for a given n and m_1 , the number of matched sets (N) required to provide a desired level of power $1 - \beta$ using a test at level α (or $\alpha/2$ two-sided) is provided by

$$N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\beta_1^2 \sigma^2 \left[\frac{m_1 m_2}{n} \right]}. \quad (7.138)$$

Hsieh and Lavori (2000) present similar expressions for a quantitative covariate in the Cox PH model.

Note that these equations involve the term $\beta_1 \sigma$ so that for a given β_1 , as the variance σ^2 increases, power increases and the required sample size decreases. The reason is that for a given β_1 , a higher value σ^2 provides a greater variation in risk over the range of covariate values. For example, over the range $\pm 3\sigma$, the odds ratio is $\exp(6\beta\sigma)$, that increases in σ .

In some cases, the effect size of interest is specified as the odds ratio per *S.D.* units (OR^σ) change in the covariate, with corresponding log odds ratio per *S.D.* of $\beta_\sigma = \log(OR^\sigma) = \beta_1 \sigma$. In this case, β_σ is substituted for $\beta_1 \sigma$ in the above expressions.

Lachin (2008) and Tang (2009) also show that power and sample size can be expressed as a function of the mean of the covariate among those with positive versus negative responses (cases vs. controls) (say μ_1 and μ_2). When m_1 and m_2 are fixed and equal for all sets, then

$$\beta_1 = \frac{\mu_1 - \mu_2}{\sigma^2}. \quad (7.139)$$

For sets with variable numbers of cases $\{m_{1i}\}$ or controls $\{m_{2i}\}$, the expression involves other constants that can be explicitly derived (see Lachin, 2008, and Tang, 2009).

7.6.7.2 Binary Covariate For a model with a single binary covariate, the expression for $I(\beta_0)$ is the same as in (7.136) with the average Bernoulli variance σ_0^2 over all sets under H_0 substituted for σ^2 . Alternatively, under H_0 it may be assumed that there is a common probability of exposure $E(X) = \pi_0$ among the cases and controls in all sets with the common Bernoulli variance $\sigma_0^2 = \pi_0(1 - \pi_0)$. Then the power of the study is provided by (7.137) and the sample size by (7.138).

7.6.7.3 Multivariate Model As described in Section 7.3.6, to assess power or sample size in a model that also adjusts for other factors, the calculations can be obtained using the variance inflation factor, that is a function of the intercorrelation among the covariates.

Example 7.20 Sample Size for a Nested Case-Control Study

Lachin (2008) describes the calculation of the sample size for a *nested case-control study* to assess the association of a quantitative biomarker with the risk of a cardiovascular event in the Diabetes Control and Complications Trial cohort. Chapter 9 describes methods for the analysis of event-time data (survival analysis) that assess the incidence of an event over time in a cohort. In a nested case-control study, a set of controls are sampled from among those still at risk (e.g., alive) at the time that an event (case) occurs. If the controls were sampled only from those who completed the study "alive," bias would be introduced because their exposure time could be far greater than that of the cases. The resulting data are then analyzed using a stratified Cox proportional hazards model that is equivalent to the conditional logistic regression model. Using (7.138), a sample size of $N = 119$ CVD cases ($m_1 = 1$ per set) matched to $m_2 = 2$ controls per case, randomly selected from among those at risk at each event time, will provide 85% power to detect an odds ratio of 1.4 per *S.D.* difference (OR^σ) using a score test at the 0.05 level, two-sided.

One of the biomarkers of primary interest is the 8-isoprostane/creatinine ratio with a standard deviation of $\sigma = 8.41$ mg/ng from preliminary data. It follows that an odds ratio of 1.4 per *S.D.* corresponds to an odds ratio of $(1.4)^{1/8.41} = 1.041$ per mg/ng difference in the biomarker, with corresponding $\beta = 0.040$. From (7.139), this in turn implies that the study would have 85% power to detect a mean difference between cases and controls of $\mu_1 - \mu_2 = \beta\sigma^2 = 3.36$ ng/mg.

Other risk factors will also be assessed. One of the most important is the presence of nephropathy. Assume that the probability of CVD is doubled from $\pi_1 = 0.1$ to $\pi_2 = 0.2$ comparing those without versus those with nephropathy, with a log odds ratio of $\beta = -0.811$. Under H_0 the assumed probability is $\pi_0 = 0.15$ that yields $\sigma_0 = 0.357$. Then the number of matched sets with one case matched to two controls required to provide 85% power with a two-sided test is $N = 161$. Conversely, a sample size of 125 sets provides 75% power.

The programs *condNpwrSD.sas* and *condNpwrbinary.sas* performed these computations.

7.7 MODELS FOR POLYCHOTOMOUS OR ORDINAL DATA

Now consider the case where the response variable is nominal or ordinal with C categories. Then we wish to estimate the probability that the i th subject with vector \mathbf{x}_i of p covariates will fall in the k th category ($k = 1, \dots, C$) of response, or $\pi_{ki}(\mathbf{x}_i, \boldsymbol{\theta}) = P(y_i = k \mid \mathbf{x}_i, \boldsymbol{\theta})$. For the case of a polychotomous nominal variable, the multinomial or polychotomous logistic regression model is employed, whereas for an ordinal response variable, it may be appropriate to consider a proportional odds model.

To be consistent with the default coding employed in the SAS PROC LOGISTIC, in this section the last (C th) category of response is considered the reference category. The response variable Y takes the values $1, \dots, C$. To present the models more conveniently, for the i th subject we define a vector of binary variables $\{Y_{ki}\}$ representing each of the $C - 1$ categories of response defined such as $y_{ki} = I\{y_i = k\}$, $I\{\cdot\}$ being the indicator function.

7.7.1 Multinomial Logistic Model

The multinomial logistic model can be employed with any polychotomous dependent variable regardless of whether it is nominal or ordinal in scaling. The model assumes that the odds of falling in the k th category of response, relative to the reference category is expressed as

$$\log \left[\frac{\pi_{ki}}{1 - \pi_{Ci}} \right] = \alpha_k + \mathbf{x}'_i \boldsymbol{\beta}_k, \quad k = 1, \dots, C - 1, \quad (7.140)$$

where $\boldsymbol{\beta}_k = (\beta_{k1} \dots \beta_{kp})^T$. The conditional probabilities as a function of the covariates are then obtained as

$$\begin{aligned} \pi_{ki} &= \frac{e^{\alpha_k + \mathbf{x}'_i \boldsymbol{\beta}_k}}{1 + \sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell}}, \quad k = 1, \dots, C - 1 \\ \pi_{Ci} &= \frac{1}{1 + \sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell}}. \end{aligned} \quad (7.141)$$

Thus, α_k is the log odds of observing the k th category of response relative to the reference category C for a covariate vector $\mathbf{x}_i = \mathbf{0}$, and β_{kj} is the log odds ratio of response k relative to category C per unit change in the covariate X_j .

Since the i th response is a multinomial variable $\{y_i\}$, then the likelihood is

$$L(\boldsymbol{\pi}) = \prod_{i=1}^N \pi_{1i}^{y_{1i}} \cdots \pi_{(C-1)i}^{y_{(C-1)i}} \pi_{Ci}^{y_{Ci}}, \quad (7.142)$$

that can also be expressed as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \left(\frac{e^{\alpha_1 + \mathbf{x}'_i \boldsymbol{\beta}_1}}{1 + \sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} \right)^{y_{1i}} \cdots \left(\frac{e^{\alpha_{(C-1)} + \mathbf{x}'_i \boldsymbol{\beta}_{(C-1)}}}{1 + \sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} \right)^{y_{(C-1)i}} \times \left(\frac{1}{1 + \sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} \right)^{1 - \sum_{\ell=1}^{C-1} y_{\ell i}}, \quad (7.143)$$

where $\boldsymbol{\theta}$ consists of the $(C-1)(p+1)$ covariates $[\alpha_1 \cdots \alpha_{C-1} \beta_{11} \cdots \beta_{1p} \cdots \beta_{(C-1)1} \cdots \beta_{(C-1)p}]$.

In effect the model is simultaneously fitting $C-1$ logistic regression models that assess the effects of covariates on the odds of falling in each of the $C-1$ categories relative to the reference category (the C th). If the model includes only a single binary covariate effect, then the simultaneous polychotomous model results match the separate category models. However, with a single quantitative covariate, or a multivariate model, the two approaches differ and the simultaneous model is preferred.

The log likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{k=1}^{C-1} y_{ki} (\alpha_k + \mathbf{x}'_i \boldsymbol{\beta}_k) - \sum_{i=1}^N \log \left(1 + \sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell} \right). \quad (7.144)$$

The score equations for α_k and β_{jk} ($k = 1, \dots, C-1$; $j = 1, \dots, p$) are

$$U(\boldsymbol{\theta})_{\alpha_k} = \frac{\partial \ell}{\partial \alpha_k} = \sum_i y_{ki} - \sum_i \frac{e^{\alpha_k + \mathbf{x}'_i \boldsymbol{\beta}_k}}{\sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} = \sum_i (y_{ki} - \pi_{ki}) \quad (7.145)$$

$$U(\boldsymbol{\theta})_{\beta_{kj}} = \frac{\partial \ell}{\partial \beta_{kj}} = \sum_i x_{ij} \left(y_{ki} - \frac{e^{\alpha_k + \mathbf{x}'_i \boldsymbol{\beta}_k}}{\sum_{\ell=1}^{C-1} e^{\alpha_\ell + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} \right) = \sum_i x_{ij} [y_{ki} - \pi_{ki}].$$

As for a simple logistic regression model, the score equation for α_k implies that the mean estimated probability for category k equals the proportion of observations in that category.

Likewise, the observed information matrix $\mathbf{i}(\boldsymbol{\theta})$ has elements

$$\begin{aligned} \mathbf{i}(\boldsymbol{\theta})_{\alpha_k \alpha_k} &= \sum_i \pi_{ki} (1 - \pi_{ki}) & \mathbf{i}(\boldsymbol{\theta})_{\beta_{kj}, \beta_{lj}} &= \sum_i -x_{ij}^2 \pi_{ki} \pi_{li} \\ \mathbf{i}(\boldsymbol{\theta})_{\alpha_k \alpha_\ell} &= \sum_i -\pi_{ki} \pi_{\ell i} & \mathbf{i}(\boldsymbol{\theta})_{\beta_{kj}, \beta_{lm}} &= \sum_i -x_{ij} x_{im} \pi_{ki} \pi_{li} \\ \mathbf{i}(\boldsymbol{\theta})_{\beta_{kj} \beta_{kj}} &= \sum_i x_{ij}^2 \pi_{ki} (1 - \pi_{ki}) & \mathbf{i}(\boldsymbol{\theta})_{\alpha_k, \beta_{kj}} &= \sum_i x_{ij} \pi_{ki} (1 - \pi_{ki}) \\ \mathbf{i}(\boldsymbol{\theta})_{\beta_{kj}, \beta_{km}} &= \sum_i x_{ij} x_{im} \pi_{ki} (1 - \pi_{ki}) & \mathbf{i}(\boldsymbol{\theta})_{\alpha_k, \beta_{\ell j}} &= \sum_i -x_{ij} \pi_{ki} \pi_{\ell i}. \end{aligned} \quad (7.146)$$

These equations then provide the basis for an iterative computation of the maximum likelihood estimates of the parameters, the estimated information matrix, and covariance matrix of the estimates.

This model can be fit using the SAS PROC LOGISTIC, or the older PROC CATMOD.

Example 7.21 *National Cooperative Gallstone Study: Biopsy Substudy*

The National Cooperative Gallstone Study (NCGS) evaluated a high and a low dose of the natural bile acid, chenodiol, for dissolution of cholesterol gallstones. Owing to signs of hepatotoxicity (liver toxicity) in monkeys, the FDA requested that a substudy be conducted in which subjects treated with the high (750 mg/day) or low (375 mg/day) dose of chenodiol have a liver biopsy prior to and at 24 months after treatment. A morphological evaluation was conducted independently by two pathologists using masked side-by-side evaluation of the baseline and follow-up biopsy, masked to order and treatment group. For each subject, each morphological characteristic was graded as worse, unchanged, or improved from baseline. Among the many characteristics assessed was sinusoidal congestion. Herein we present an analysis of the evaluations by one of the two pathologists. While the grading is ordinal, we start with an analysis of the polychotomous response variable WUI coded as worsening (1) or unchanged (2) relative to improved (3) as the reference category.

The data set is contained in the program *NCGS biopsy.sas* and includes the variables TRETGRP (1 high dose, 2 low dose), sex (0 female, 1 male), and age (years), among others. The following table describes the crude percentages within each response category for each treatment group

<i>Treatment</i>	<i>Worse</i>	<i>Unchanged</i>	<i>Improved</i>
High dose (1)	8 (25.8%)	3 (9.7)	20 (64.5)
Low dose (2)	8 (22.2)	9 (25.0)	19 (52.8)

The following statements fit the multinomial logistic model with treatment group as the only covariate:

```
proc logistic data = kcat;
class tretgrp / param=ref;
model wui = tretgrp / link=glogit;
```

The *link=glogit* specifies that the response function employed is the generalized logit that corresponds to the model parameterization for the multinomial logistic model. The *param=ref* specifies that the class variable *tretgrp* enter into the design matrix as a binary variable relative to the reference category that is the highest value among those in the data, in this case the high- versus low-dose group. The $\hat{\beta}_k$ equals the log odds ratio of falling in category *k* relative to category *C* per unit change in the class covariate *X* (i.e., comparing one treatment group to another with binary reference coding). The $-2 \log L = 129.310$ for the no-intercept model and the model likelihood ratio chi-square test is 2.79 on 2 *df* with *p* = 0.2476. The Wald test of the treatment group effect yields *p* = 0.2868 on 2 *df*. The treatment group effect is tested on 2 *df* because the model estimates two effects, one for the comparison of worsening versus improved, the other for unchanged versus improved. The respective estimates provided by the program are

Analysis of Maximum Likelihood Estimates

Parameter	WUI	DF	Estimate	Standard Error	Wald	Pr >
					Chi-Square	ChiSq
Intercept	1	1	-0.1178	0.4859	0.0588	0.8085
Intercept	2	1	0.7472	0.4047	3.4098	0.0648
TRETGRP	1	1	1.0986	0.8333	1.7379	0.1874
TRETGRP	1	2	1.1499	0.7396	2.4169	0.1200

and the resulting odds ratios for the two treatment effect estimates are

Effect	WUI	Odds Ratio Estimates	
		Odds Ratio	95% Confidence Limits
TRETGRP 1 vs 2	1	3.000	0.586 15.362
TRETGRP 1 vs 2	2	3.158	0.741 13.458

As described above, a polychotomous logistic model in effect fits two separate models simultaneously that represent the covariate effects (treatment group in this case) on the odds of the k th category (1 or 2) versus the reference (3). In a simple logistic model with a binary response, as shown in Chapter 6, the intercept (α) is the logit of the probability within the reference covariate category ($tretgrp = 2$ in this case) and the coefficient (β) is the log odds ratio for the other group versus the reference. The description above is a simple generalization. The parameter estimates in order are α_1 , α_2 , β_1 , and β_2 . The odds ratio of 3.0 for treatment group 1 versus group 2 for $WUI = 1$ is the odds ratio for worsening ($WUI = 1$) versus the reference improved ($WUI = 3$); and the value 3.158 for $WUI = 2$ is the odds ratio for unchanged versus improved. Because the data consist of two sets of three proportions, each adding to 1, then this is a "saturated" model on 4 df and the resulting estimates of the logits yield estimated probabilities equal to the observed proportions above.

The following statements fit the model with treatment group adjusted for sex and age:

```
proc logistic;
class tretgrp sex / param=ref;
model wui = tretgrp sex age / link=glogit;
```

The results are presented in Table 7.12. The model likelihood ratio chi-square test value is 13.07 on 6 df with $p = 0.0419$. For each covariate effect a 2 df Wald type 3 test is computed followed by the coefficient estimates of each effect on the logits of each category of response versus the reference category. The respective odds ratios are also shown. None of the 2 df tests of effects are nominally significant. However, the 1 df test of the sex effect on the logit of worsening ($WUI = 1$) versus the reference (improvement) is nominally significant ($p = 0.0225$) with a corresponding odds ratio of 0.121. Since *sex* enters as a class effect with categories (0, 1), the reference

Table 7.12 Multinomial logistic regression analysis of NCGS biopsy gradings of worse (1), unchanged (2), or improved (3).

The LOGISTIC Procedure							
Type 3 Analysis of Effects							
Effect	DF	Chi-Square		Pr > ChiSq			
		Wald					
TRETGRP	2	3.0292		0.2199			
SEX	2	5.2937		0.0709			
AGE	2	4.7603		0.0925			
Analysis of Maximum Likelihood Estimates							
Parameter	WUI	DF	Estimate	Error	Chi-Square	Pr > ChiSq	
Intercept	1	1	5.2609	2.8755	3.3472	0.0673	
Intercept	2	1	1.3734	2.5156	0.2981	0.5851	
TRETGRP	1	1	1.4489	0.8920	2.6384	0.1043	
TRETGRP	1	2	1.2248	0.7663	2.5545	0.1100	
SEX	0	1	1	-2.1111	0.9254	5.2045	0.0225
SEX	0	2	1	-1.4102	0.7822	3.2507	0.0714
AGE		1		-0.0843	0.0521	2.6223	0.1054
AGE		2		0.00354	0.0436	0.0066	0.9353
Odds Ratio Estimates							
Effect	WUI	Estimate	Point		95% Wald		
			Confidence	Limits			
TRETGRP	1 vs 2	1	4.258	0.741	24.462		
TRETGRP	1 vs 2	2	3.404	0.758	15.283		
SEX	0 vs 1	1	0.121	0.020	0.743		
SEX	0 vs 1	2	0.244	0.053	1.131		
AGE		1	0.919	0.830	1.018		
AGE		2	1.004	0.921	1.093		

category is 1 (male) so that the odds ratio represents the female versus male odds ratio, the odds for females being 12.1% that of males.

7.7.2 Proportional Odds Model

Clayton (1974), among others, explored the use of a logit model with cumulative logits from an ordinal response variable within separate groups. McCullagh (1980) described the proportional odds model that allowed the assessment of covariate effects on the cumulative logits. McCullagh and Nelder (1989) describe this and other models for an ordinal response variable. The motivation was a model that would assess the effect of covariates on the levels of a quantitative response variable that had been grouped or categorized using cut points. However, the assumption of an underlying quantitative response variable is not necessary to apply the model.

The cumulative logits were described in Section 2.10, but not explicitly. For the j th category, let $\pi_{j+} = \sum_{\ell=1}^j \pi_j$ denote the cumulative probability up to and including the j th category and let the corresponding odds be designated as

$$\kappa_j = \frac{\pi_{j+}}{1 - \pi_{j+}}. \quad (7.147)$$

For the i th subject with covariate vector \mathbf{x}_i , the model specifies that

$$\log(\kappa_{ij}) = \alpha_j + \mathbf{x}'_i \boldsymbol{\beta}, \quad j = 1, \dots, C, \quad (7.148)$$

with parameter vector $\boldsymbol{\theta} = (\alpha_1 \ \dots \ \alpha_C \ \beta_1 \ \dots \ \beta_p)^T$. This model provides proportional odds because any two subjects with covariate vectors \mathbf{x}_1 and \mathbf{x}_2 then have odds that are proportional, or

$$\frac{\kappa_{1j}}{\kappa_{2j}} = \exp[(\mathbf{x}_1 - \mathbf{x}_2)' \boldsymbol{\beta}] \quad (7.149)$$

for all categories $j = 1, \dots, C$. Thus, the α_j is the log cumulative odds for category j for a subject with covariate vector $\mathbf{x}_i = 0$. If all covariates are mean centered (i.e., have mean zero) before fitting the model, then α_j is the mean log cumulative odds in the population, and $\hat{\alpha}_j$ that in the sample, for the j th category.

Since π_{j+} is a function of the C category probabilities, this implies that the multinomial likelihood function can be expressed as a function of the π_{j+} . To simplify notation, let $\gamma_j = \pi_{j+}$, where $\gamma_1 = \pi_1$. Consider a model without covariates where the numbers observed in the C categories are designated as n_j with cumulative sums $R_j = \sum_{\ell=1}^j n_{\ell}$, $N = R_C$, and corresponding proportion $Z_j = R_j/N$. McCullagh (1980) shows that the multinomial likelihood equals

$$L = \prod_{j=1}^C \pi_j^{n_j} = \left(\frac{\gamma_1}{\gamma_2} \right)^{R_1} \left(\frac{\gamma_2 - \gamma_1}{\gamma_2} \right)^{R_2 - R_1} \times \left(\frac{\gamma_2}{\gamma_3} \right)^{R_2} \left(\frac{\gamma_3 - \gamma_2}{\gamma_3} \right)^{R_3 - R_2} \dots \left(\frac{\gamma_{C-1}}{\gamma_C} \right)^{R_{C-1}} \left(\frac{\gamma_C - \gamma_{C-1}}{\gamma_C} \right)^{R_C - R_{C-1}}. \quad (7.150)$$

This then generalizes to a model with subject-specific probabilities as a function of covariate values. The model is then fit by applying the principles of maximum likelihood estimation.

The proportional odds model is a reduced model compared to a cumulative logit model with heterogeneous odds, or a nonproportional odds model, of the form

$$\log(\kappa_{ij}) = \alpha_j + \mathbf{x}'_i \boldsymbol{\beta}_j, \quad j = 1, \dots, C. \quad (7.151)$$

While PROC LOGISTIC does not fit this more general model, it does provide a score test of the simplifying assumption, or of the hypothesis $H_0: \boldsymbol{\beta}_j = \boldsymbol{\beta} \forall j$ (see Peterson and Harrell, 1990).

Example 7.22 *National Cooperative Gallstone Study: Biopsy Sub-study (continued)*

The following statements fit the proportional odds model for treatment group adjusted for sex and age:

```
proc logistic;
class tretgrp sex / param=ref;
model wui = tretgrp sex age;
```

The likelihood ratio chi-square test value is 8.92 on 3 *df* with $p = 0.0303$. The score test value for the omitted effects in the nonproportional odds model in (7.151) with $C - 1$ coefficients per covariate, versus the proportional odds model in (7.148) with a single coefficient per covariate, is 5.311 on 3 *df* with $p = 0.1504$. Thus, the nonproportional odds model would not provide a better fit for this ordinal response variable.

The estimates provided by the program are

Parameter	DF	Analysis of Maximum Likelihood Estimates			
		Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept 1	1	2.2524	1.6990	1.7575	0.1849
Intercept 2	1	5.2534	1.8273	8.2657	0.0040
TRETGRP	1	0.7402	0.5025	2.1702	0.1407
SEX	0	-1.2785	0.5329	5.7565	0.0164
AGE	1	-0.0617	0.0311	3.9458	0.0470

The two intercepts are substantially different from those of the multinomial logit model presented in Table 7.12. The latter represent the logit of the probability of worsening for an observation with other covariate values zero, whereas those in the proportional odds model represent the logit of the cumulative probability; i.e., *Intercept 1* = $\text{logit}(\pi_{1+})$ and *Intercept 2* = $\text{logit}(\pi_{2+})$ for covariate vector $\mathbf{x} = 0$.

There is a single coefficient for each covariate with the corresponding odds ratios as follows:

Effect	Odds Ratio Estimates		
	Odds Ratio	95% Wald Confidence Limits	
TRETGRP 1 vs 2	2.096	0.783	5.613
SEX 0 vs 1	0.278	0.098	0.791
AGE	0.940	0.885	0.999

Each is the odds ratio for the cumulative logit of category 1 versus (2,3) and that of categories (1,2) versus 3. The odds ratio for treatment group is not significant whereas those for sex and age are now nominally significant, the odds being lower among females and decreasing with increasing age.

7.7.3 Conditional Models for Matched Sets

In principle the conditional binary logistic model could be generalized to polychotomous responses. However, these models are not available through SAS and are not described herein.

7.8 RANDOM EFFECTS AND MIXED MODELS

A thorough treatment of mixed models for categorical outcomes is beyond the scope of this book. Those interested are referred to books on nonlinear mixed models, such as Demidenko (2004). Herein we only consider simple mixed models as would be applied in a meta analysis or similar analysis adjusting for a random effect.

The concept of a random effects model was introduced in Section 4.10 in conjunction with a meta-analysis of multiple studies. Following the work of DerSimonian and Laird (1986), various authors proposed models with a random effect for the parameter of interest (e.g., log odds ratio) among studies. The book by Whitehead (2002) provides a review of these methods and shows how these models can be fit using a mixed effects model program, such as SAS PROC MIXED. Herein we take a different approach by modeling the probability of response as a function of fixed and random effects.

7.8.1 Random Intercept Model

As shown in Section 7.1.4, a stratified analysis of K 2×2 tables of group by response can be performed using additional binary variables to represent the additional effects of strata in a logistic model. Alternatively, a variable for strata could be specified in a *class* statement and included in the *model* statement. In either approach, strata enter the model as a fixed effect with $K - 1$ coefficients representing a contrast of each stratum with the reference stratum (using reference coding).

In some cases, such as a meta analysis over a set of studies, or an analysis adjusted for clinic or site effects, the stratifying variable can be considered to represent a

"random sample" from a population, or an effect that for whatever reason varies randomly in the population. To simplify, consider a model with a single covariate (say treatment group, $X = 0$ or 1) and a random stratum variable. The effect of strata can then be represented by a random intercept model such as

$$\text{logit}(\pi_{ij}) = \bar{\alpha} + \alpha_i + x_{ij}\beta \quad (7.152)$$

for the j th subject in the i th stratum, where it is assumed that the random stratum effects $\{\alpha_i\}$ are in turn normally distributed with mean 0 and variance σ_α^2 about the overall mean intercept $\bar{\alpha}$. This then provides the following expressions for the logits of the probabilities within the two groups among strata

Stratum	$X = 0$	$X = 1$
1	α_1	$\alpha_1 + \beta$
2	α_2	$\alpha_2 + \beta$
\vdots	\vdots	\vdots
K	α_K	$\alpha_K + \beta$

(7.153)

Thus, β represents the common odds ratio and the intercepts represent the random stratum effects with mean $\bar{\alpha}$.

The model has two fixed effect parameters, the $\bar{\alpha}$ and β . In addition, the model can provide estimates of the random effect for each stratum; i.e., the $\{\alpha_i\}$. The model also provides an estimate of the random effect variance σ_α^2 and its S.E. that can be employed to assess the appropriateness of a random versus fixed effects model.

For the above structure, let n_i denote the number of observations within the i th stratum with total sample size N . With a normally distributed random effect in a logistic regression model, generalizing to a covariate vector x_{ij} with coefficient vector β for the fixed effects, the predicted value of the probability for a subject in the i th stratum is

$$\int_{-\infty}^{\infty} \pi_{ij} d\Phi \left[\frac{\alpha_i - \bar{\alpha}}{\sigma_\alpha} \right], \quad (7.154)$$

where $d\Phi(\cdot)$ represents the standard normal density. Then the pseudo-likelihood function is

$$\left[\sqrt{2\pi}\sigma \right]^{-N} \prod_{ij} \int_{-\infty}^{\infty} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \exp \left[\frac{-(\alpha_i - \bar{\alpha})^2}{2\sigma_\alpha^2} \right] d\alpha_i, \quad (7.155)$$

and the log likelihood ℓ is proportional to

$$\sum_{ij} \log \int_{-\infty}^{\infty} \left\{ \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \exp \left[\frac{-(\alpha_i - \bar{\alpha})^2}{2\sigma_\alpha^2} \right] \right\} d\alpha_i, \quad (7.156)$$

where the π_{ij} are expressed as a logistic function of the linear predictor in (7.152). The estimates of the parameters $\bar{\alpha}$, σ_α^2 , and β are then obtained by maximizing the restricted pseudo-likelihood. See Demidenko (2004) and the online SAS PROC GLIMMIX documentation for computational details.

7.8.2 Random Treatment Effect

A generalization would then be to also assume that the treatment group effect varies at random among strata, as in the case where a test of homogeneity would indicate significant heterogeneity among strata. The model then becomes

$$\text{logit}(\pi_{ij}) = \bar{\alpha} + \alpha_i + x_{ij}(\bar{\beta} + \beta_i). \quad (7.157)$$

The expected logits are then obtained as in the table above using the stratum-specific $\{\alpha_i, \beta_i\}$ that are assumed to be jointly distributed as bivariate normal with covariance matrix G with components σ_α^2 , σ_β^2 , and $\sigma_{\alpha\beta}$.

Example 7.23 DCCT Treatment Group Adjusted for Clinic (continued)

Example 7.17 describes the assessment of the effects of intensive versus conventional treatment on the risk of microalbuminuria when adjusting for differences among clinical centers using added fixed clinic effects or by conditioning on clinic. An alternative approach would be to add clinic to a logistic model as a normally distributed random effect. This would be obtained using the following statements in the program *ClinRenal.sas*:

```
proc glimmix data = renal;
  class inten female clinic;
  model micro24 (descending) = inten hbael yearsdm sbp female
    / solution dist=binary link=logit;
  lsmeans inten / cl ilink;
  random intercept / subject = clinic;
```

The variable *int* to designate intensive versus conventional treatment has been renamed *inten* because in the *random* statement *int* is recognized as *intercept*.

In order to model the probability that *micro24* = 1 rather than 0, the *descending* option is used, but in a different syntax from that in the other procedures. Table 7.13 presents the results. The *generalized chi-square* is a generalization of the Pearson goodness-of-fit statistic to allow for correlated observations (e.g., repeated measures). As with PROC GENMOD, a *value/df* close to 1.0 is desirable.

The *random intercept* statement uses "subject =" to specify *clinic* as a random (intercept) effect with mean zero. The fixed intercept $\bar{\alpha} = -8.0180$ is the mean of the within-clinic random intercepts. The model-estimated variance is $\hat{\sigma}_\alpha^2 = 0.04780$ with *S.E.* = 0.1053. Clearly, this variance component is not significantly different from zero and the random intercept might be dropped from the model.

For class effects, such as *inten*, the higher-ordered category is used as the reference. Thus, the coefficient in this model for *inten* is the log odds ratio for conventional versus intensive therapy, whereas previous models used *inten* = 0 as the reference to model the log odds ratio of intensive versus conventional therapy. Thus, $-\hat{\beta} = -0.6131$ for *inten* in Table 7.13 would be compared to those from prior analyses. With this conversion, the coefficients of the fixed effects, their *S.E.*, and *p*-values are very close to those in a simple logistic model not adjusted for clinic, as in Section 7.6.4.

Table 7.13 Mixed logistic regression model of DCCT microalbuminuria data with a random clinic effect.

The GLIMMIX Procedure						
Fit Statistics						
-2 Res Log Pseudo-Likelihood						2706.84
Generalized Chi-Square						511.27
Gener. Chi-Square / DF						0.96
Covariance Parameter Estimates						
Standard						
Cov Parm	Subject	Estimate		Error		
Intercept	CLINIC	0.04780		0.1053		
Solutions for Fixed Effects						
Standard Pr > t						
Effect	Estimate	Error	DF	t Value	t	Pr > t
Intercept	-8.0180	1.6452	20	-4.87	<.0001	
INT 0	0.6131	0.2474	512	2.48	0.0135	
INT 1	0	
HBAEL	0.4186	0.0797	512	5.25	<.0001	
YEARSMD	0.0978	0.0362	512	2.70	0.0072	
SBP	0.0078	0.0117	512	0.67	0.5017	
FEMALE 0	0.5851	0.2733	512	2.14	0.0327	
FEMALE 1	0	
INT Least Squares Means						
Standard						
INT	Estimate	Error	DF	t Value	Pr > t	Alpha
0	-1.5187	0.1758	512	-8.64	<.0001	0.05
1	-2.1318	0.2048	512	-10.41	<.0001	0.05
Standard						
Error Lower Upper						
INT	Lower	Upper	Mean	Mean	Mean	Mean
0	-1.8640	-1.1734	0.1797	0.02590	0.1342	0.2362
1	-2.5341	-1.7295	0.1060	0.01941	0.07350	0.1507

Since the likelihood is based on a normally distributed random effect, the test statistics are t (or F)-tests. The denominator degrees of freedom for the fixed covariate effects are 512 that equals the total N minus the number of random and fixed effects (538 – 21 – 5).

The *lsmeans* produced by the model are the adjusted mean logits. The *mlink* option also requests that the inverse link be applied so that in this case the lsmean probabilities within each treatment group and their confidence limits are computed.

The model could be expanded to include both a random intercept and a random treatment group effect using the statement

```
random intercept inten / subject = clinic;
```

Although the solution of the model fixed effects converges, the SASLOG includes the warning "Estimated G matrix is not positive definite." The reason is that the model is unable to jointly estimate the variance components of the random effects, the estimate for the variance among *clinic* random effects being zero. Thus, the random intercept is unnecessary and can be dropped from the model.

The model could be refit with only a random treatment group effect among clinics using the statement

```
random inten / subject = clinic;
```

in addition to the other statements above. The covariate *clinic* is used only to define the blocks over which the random *inten* effect is computed; i.e., *clinic* does not appear in the *model* statement. The resulting variance component estimate is $\hat{\sigma}_{\beta}^2 = 0.1348$ with *S.E.* = 0.1641, that also is not statistically significant. The estimate for the fixed *inten* effect is then the estimate of the mean effect $\bar{\beta}$ over blocks (clinics) that equals 0.6267 with *S.E.* = 0.2738 and *p* = 0.0275, similar to that in Table 7.13 and the other models.

7.9 MODELS FOR MULTIVARIATE OR REPEATED MEASURES

Section A.10 describes the family of quasi-likelihood generalized linear regression models (*GLMs*) for members of the exponential family of distributions, and A.11 describes the family of like models for multivariate or repeated measurements using generalized estimating equations (*GEE*). The latter then provides models that can be applied to multivariate responses or to repeated measures. These models are also fit using the SAS procedure GENMOD. We first consider the analysis of a set of repeated measures followed by a multivariate analysis. We do so principally by computational examples.

7.9.1 GEE Repeated Measures Models

Let y_{ij} denote the binary response observed for the i th subject (or unit or cluster $i = 1, \dots, N$) at the j th repeated assessment ($j = 1, \dots, K$), where some of the replicates may be missing at random (see Section A.11). Associated with each replicate of each subject is a covariate vector \mathbf{x}_{ij} that includes fixed (or baseline) covariates that are characteristics of the subject, and time-dependent covariates that are associated with the j th replicate within subjects. The logistic model then specifies that the logit of the probability of a positive response, π_{ij} , is a linear function of the covariates of the now-familiar form

$$\log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \mathbf{x}'_{ij} \boldsymbol{\theta}. \quad (7.158)$$

In order to fit the model, it is necessary to specify the working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ among the K replicate measures for each subject, expressed as a function of a vector of correlation parameters $\boldsymbol{\alpha}$. With large samples, the *GEE* estimates $\hat{\boldsymbol{\theta}}$ are consistent estimates of $\boldsymbol{\theta}$ regardless of whether the working correlation matrix equals the actual correlation matrix. Some options are the independence correlation structure with $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}_K$, the identity matrix; or the exchangeable correlation matrix where $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I} + \boldsymbol{\alpha} \mathbf{1} \mathbf{1}'$, $\mathbf{1}$ being the ones vector; or the first-order autoregressive AR(1) model where $\mathbf{R}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{|j-k|}$ for $j \neq k$; and the unstructured matrix where all $K(K - 1)/2$ correlations are estimated from the model. The model is then fit using the generalized quasi-likelihood estimating equation, and the covariance matrix of the parameter estimates is estimated using a generalization of the information sandwich robust estimate (see Section A.11).

As with a logistic regression model, analysis of longitudinal or clustered data with *GEE* is valid with large samples. With *GEE*, however, the number of subjects or clusters required for valid inferences is greater than that for logistic regression due to the need to estimate the within-subject covariance structure. Lu, et al. (2007) present a review of the considerations in the application of the sandwich estimator in *GEE* and a comparison of some of the corrections that may improve the small sample properties.

While *GEE* provides consistent estimates of the model coefficients with a possibly incorrect working correlation structure, the closer the assumed correlation structure is to the true (unknown) structure, the more efficient are the coefficient estimates. Thus, if the sample size is adequate, the unstructured correlation matrix is preferred; i.e., that empirically estimated from the data.

Once fit, a variety of statistical tests can be conducted. Section 4.5 describes the partitioning of the general omnibus multivariate test for elements of a vector of K parameters into a test of homogeneity among the parameters and an overall test of association. These tests also apply to the analysis of a set of parameters related to the analysis of repeated measures. This is illustrated in the following example.

Example 7.24 DCCT Nephropathy Data (continued)

Example 7.4 presents the prevalence of microalbuminuria after six years of follow-up in a sample of 172 subjects in the secondary intervention cohort with a high normal

albumin excretion rate (*AER*) at baseline defined as $15 \leq AER \leq 40$. In fact, a total of 225 such patients were entered into the study of whom 220 were evaluated during years 4 to 7 of annual follow-up: 218 at year 4, 216 at year 5, 172 at year 6, and 124 at year 7 of follow-up. Virtually all of the missing data is attributable to administrative curtailment of follow-up as a function of when the subject entered the study, and thus satisfies the assumption that the missing data are missing completely at random, in the sense defined by Little and Rubin (2002). The following table presents the prevalence of microalbuminuria at each year (%) among those evaluated (*n*) within the intensive and conventional treatment groups.

	4	5	6	7
<i>Intensive % (n)</i>	8.3 (109)	13.9 (108)	12.4 (89)	11.9 (67)
<i>Conventional % (n)</i>	27.5 (109)	31.5 (108)	37.4 (83)	40.4 (57)

To conduct a longitudinal or repeated measures analysis of these data using *GEE*, the data set must be constructed with one record per subject per replicate measurement. For these data, for example, the following are the data for two subjects

patient	year	micro	inten	SBP
156	4	0	0	110
156	5	0	0	110
156	6	1	0	110
156	7	0	0	110
932	4	0	1	114
932	5	0	1	114

The variable *year* is the year of the repeated measurement, *inten* the indicator variable for intensive (1) versus conventional (0) treatment, *micro* the response value for that repeated measure, and *sbp* is the baseline systolic blood pressure. The values of *inten* and *sbp* are fixed and do not vary over replicate measures.

The following statements would then fit a *GEE* logit model to assess the unadjusted treatment group effects at each year of follow-up, assuming an independence working covariance structure:

```
proc genmod data = renlong descending;
class patient inten year / descending;
model micro = year inten(year) / dist = binomial link = logit;
lsmeans inten(year);
repeated subject = patient / type=independent;
```

This model computes a coefficient for *inten(year)* that is the log odds ratio of intensive versus conventional therapy at each year of follow-up. The *lsmeans* for the *inten(year)* effect provides estimates of the proportions within each group at each year that match perfectly the crude percentages given above.

An unstructured covariance structure would be provided by

```
repeated subject = patient / type=unstructured corrw;
```

In this case, the *lsmeans* would not match the empirical proportions. Rather, the estimates, expressed as a percent, are

	4	5	6	7
Intensive %	8.16	13.7	13.4	11.5
Conventional %	27.4	31.7	33.5	34.9

Since little data are missing at years four and five (two and four observations, respectively), the *GEE-unstructured* model-estimated percentages are almost identical to the crude percentages; but at years six and seven, where more data are missing, the model estimates are somewhat different, especially in the conventional group at year seven. The reason is that the model is now taking into account the intercorrelations among the observed values to estimate the marginal mean (probability) that would have been observed had there not been any missing data. Thus, the model using the unstructured covariance matrix would be preferred.

The *corrw* option provides the empirical correlation matrix among the replicates.

Working Correlation Matrix				
	Col1	Col2	Col3	Col4
Row1	1.0000	0.6673	0.5597	0.7133
Row2	0.6673	1.0000	0.6554	0.7843
Row3	0.5597	0.6554	1.0000	0.8176
Row4	0.7133	0.7843	0.8176	1.0000

It is clear that the pattern does not strongly satisfy one of the other possible working correlation structures with fewer parameters. Thus, we will use this unstructured correlation approach for the remaining models.

The following are the model estimates provided:

Parameter	Estimate	Error	Limits	Z	Pr > Z
Intercept	-0.9764	0.2144	-1.3966 -0.5562	-4.55	<.0001
YEAR 7	0.3546	0.1884	-0.0146 0.7238	1.88	0.0598
YEAR 6	0.2902	0.1878	-0.0779 0.6583	1.55	0.1223
YEAR 5	0.2099	0.1473	-0.0787 0.4986	1.43	0.1540
YEAR 4	0.0000	0.0000	0.0000 0.0000	.	.
inten(YEAR) 1 7	-1.4235	0.4154	-2.2378 -0.6092	-3.43	0.0006
inten(YEAR) 0 7	0.0000	0.0000	0.0000 0.0000	.	.
inten(YEAR) 1 6	-1.1810	0.3699	-1.9059 -0.4560	-3.19	0.0014
inten(YEAR) 0 6	0.0000	0.0000	0.0000 0.0000	.	.
inten(YEAR) 1 5	-1.0727	0.3464	-1.7516 -0.3938	-3.10	0.0020
inten(YEAR) 0 5	0.0000	0.0000	0.0000 0.0000	.	.
inten(YEAR) 1 4	-1.4438	0.4107	-2.2487 -0.6390	-3.52	0.0004
inten(YEAR) 0 4	0.0000	0.0000	0.0000 0.0000	.	.

An equivalent model could be fit with an interaction effect, as described subsequently. However, the nested model has the advantage that it provides an estimate of the treatment group effect separately by year, with a test of the significance of that effect. Let $(\beta_4, \beta_5, \beta_6, \beta_7)$ denote the log odds ratio between the treatment groups at years four to seven, respectively. The estimates are the values above for each *inten(year)* effect. Each is highly statistically significant nominally (with no adjustment for multiple tests).

Section 4.5 describes multivariate tests of hypotheses and the partitioning of the omnibus test into a test of homogeneity and a test of association. In the analysis of repeated measurements, the omnibus test is called the MANOVA test since it is analogous to a *multivariate analysis of variance* test for repeated measures. In the present example, the omnibus or MANOVA test then provides a test of the multivariate null hypothesis $H_0: \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ against the omnibus alternative $H_{1O}: \beta_j \neq 0$ for some $j \in (4, \dots, 7)$. Using the above nested model, the MANOVA omnibus test is provided by the overall type 3 test of the *inten(year)* effect. The resulting omnibus test value is $X_O^2 = 16.34$ that is distributed as chi-square on 4 *df*, with $p = 0.0026$.

However, as was the case with an analysis of stratified 2×2 tables, rather than testing that all log odds ratios are equal to 0, versus one or more differing from 0, it may be more meaningful to describe the average log odds ratio over the repeated measures, and test the hypothesis of homogeneity.

The hypothesis of homogeneity specifies $H_{0H}: \beta_4 = \beta_5 = \beta_6 = \beta_7$ versus $H_{1H}: \beta_j \neq \beta_k$ for $4 \leq j < k \leq 7$. This is tested by a model with the *inten*year* interaction as follows:

```
proc genmod data = renlong descending;
class patient inten year / descending;
model micro = inten year inten*year
  / dist = binomial link = logit type3;
repeated subject = patient / type=unstructured;
```

This yields a type 3 test of interaction with $X_H^2 = 1.87$ on 3 *df*, and $p = 0.5988$ that is clearly not statistically significant.

The hypothesis of overall association then specifies $H_{0A}: \beta_j = \beta = 0, \forall j \in (4, \dots, 7)$ against the alternative $H_{1A}: \beta \neq 0$. The estimate of the average group effect over all evaluations and the test of association are provided by a simple no-interaction model using the statements

```
class patient inten year / descending;
model micro = inten year / dist = binomial link = logit type3;
lsmeans inten;
```

The overall test of the intensive group difference over the four visits yields $X_A^2 = 15.71$ on 1 *df* with $p < 0.0001$. The estimated log odds ratio for intensive versus conventional treatment is $\hat{\beta} = -1.210$ with empirical (robust) *S.E.* = 0.313 and 95%

confidence limits of $(-1.823, -0.598)$. The corresponding odds ratio is 0.298 with confidence limits $(0.161, 550)$. The estimated LSMEANS and confidence limits are 0.120 (0.077, 0.182) in the intensive group and 0.314 (240, 0.399) in the conventional group. The confidence limits are obtained from the logistic function of the confidence limits of the L' Beta values as described in Example 7.7.

Finally, as in a simple logistic regression model, additional qualitative or quantitative covariate effects could be added to the *model* statement. These effects could be a mixture of fixed (e.g., baseline) covariates that apply to all repeated measurements, or time-dependent covariates that are associated with each particular repeated measurement. For example, rather than assessing the effects of baseline systolic blood pressure on the risk of microalbuminuria, we could assess the effects of blood pressure as a time-dependent covariate. In this case, the data set would be constructed as above but using the value of *sbp* observed at each follow-up year.

In some cases, the repeated measures may represent within-subjects design effects. For example, suppose that biopsy specimens are obtained at six and 12 months of follow-up and each biopsy is read by two pathologists: A (1) and B (0). Then each subject can contribute up to four replicate observations within the time*reader design. Thus, the data for a given subject might appear as

patient	response	month	reader
12	0	6	0
12	1	12	0
12	0	6	1
12	0	12	1

This would also apply to cluster designs where the *repeated* statement would use the cluster ID in place of the patient ID in the *subject* = specification. Then a variable for member would be used in place of time or month.

These computations are provided by the program *renallong09.sas*.

7.9.2 GEE Multivariate Models

GEE can also be applied to multivariate observations where multiple, but different, measurements are obtained from each subject. For example, suppose that we wished to assess the effect of intensive treatment on the risk of retinopathy and nephropathy, jointly, averaged over five and six years of follow-up. Then the data set would contain a variable, say *measure*, coded as 1 if retinopathy and 2 if nephropathy. Then the data set would be constructed as follows:

patient	measure	response	inten	year
47	0	0	0	5
47	1	1	0	6
47	0	0	0	5
47	1	1	0	6

Then the model would be fit using the statements

```

class patient inten measure year / descending;
model response = measure year inten(measure)
  / dist = binomial link = logit;
repeated subject = patient / type=unstructured;

```

The test of the *inten(measure)* effect then provides a 2 *df* test of the treatment group effect on either measure averaged over five and six years.

7.9.3 Random Coefficient Models

Models such as the above are considered repeated measures models. An alternative method to capture the dependence among repeated measures is to assume a random effects model. For the basic analysis above with repeated assessments of nephropathy at 4 to 7 years of follow-up, an alternative approach is to describe the mean rate of change in the odds of nephropathy as a linear function of the year of the follow-up assessment. This could be assessed using PROC GLIMMIX.

The estimate of the rate of change within each group would be provided by a nested model such as the following:

```

proc glimmix data = renlong;
class inten female;
model micro (descending) = inten year(inten)
  / solution dist=binary link=logit;
random year / subject = patient;

```

Note that *year* enters the model as a linear quantitative effect. The model assumes that the subject-specific coefficients for the *year* effect then are normally distributed with an overall mean within each treatment group. The *year(inten = 0)* estimate from the model is 0.1532. This is the mean change in the log odds of prevalence of nephropathy per year of follow-up in the conventional treatment group, or an odds ratio of 1.166 per year, or a 17% greater odds per year, on average. This rate of change, however, is not significantly different from zero. Likewise, the mean change in the log odds per year in the intensive group is provided by the *year(inten = 1)* coefficient estimate of -0.00218 that corresponds to an odds ratio of 0.998 per year, or virtually no change that is also not significantly different.

A test of the equality of these two average rates of change per year is provided by an additional model using an interaction in the *model* statement such as the following:

```
model micro (descending) = inten year inten*year / .....
```

The test of the difference between the coefficients in each group is provided by the type 3 test of the interaction effect that is not significant at $p = 0.537$.

In conclusion, the *GEE* repeated measures analysis shows that over the period of follow-up of 4 to 7 years, the prevalence of microalbuminuria was on average significantly higher in the conventional group than in the intensive treatment group.

Table 7.14 Allergic reactions in lumbar surgery as a function of route of administration and gender, from Thisted (1988), reproduced with permission.

Category <i>j</i>	# Allergic Reactions <i>a_j</i>	# Surgeries <i>n_j</i>	(Intercept, Route, Gender) (1, <i>x_{1j}</i> , <i>x_{2j}</i>)
1	76	20,959	(1, 0, 0)
2	132	12,937	(1, 0, 1)
3	15	6,826	(1, 1, 0)
4	29	4,191	(1, 1, 1)

However, the *GLIMMIX* model analysis shows that the mean rate of change in the prevalence per year did not differ significantly between groups.

7.10 PROBLEMS

7.1 Consider a logistic regression model with two covariates and $\boldsymbol{\theta} = (\alpha \ \beta_1 \ \beta_2)^T$. Show that:

- 7.1.1. The likelihood is as expressed in (7.4).
- 7.1.2. $\log L(\boldsymbol{\theta})$ is as expressed in (7.5).
- 7.1.3. $U(\boldsymbol{\theta})_\alpha = m_1 - \sum_i \pi_i$.
- 7.1.4. $U(\boldsymbol{\theta})_{\beta_j} = \sum_i x_{ij}(y_i - \pi_i)$ for $j = 1, 2$.
- 7.1.5. The observed information matrix $\mathbf{i}(\boldsymbol{\theta})$ has elements as in (7.11)–(7.14).

7.2 Use a logistic regression model to conduct an adjusted analysis over strata for the stratified 2×2 tables in Chapter 4. Compare the logistic model estimated *MLEs* of the odds ratios, their asymmetric confidence limits, the likelihood ratio, and Wald tests to the results obtained previously using the *MVLE* and Mantel-Haenszel procedures for:

7.2.1. The ulcer clinical trial data presented in Example 4.1 and with results shown in Example 7.3.

7.2.2. The religion and mortality data of Example 4.6.

7.2.3. The pre-eclampsia data of Example 4.24.

7.2.4. The smoking and byssinosis data of Problem 4.9.

7.3 Thisted (1988, p.194) presents the results of a study conducted to assess the influence of the route of administration (X_2 : 1 = local, 0 = general) of the anesthetic chymopapain and gender (X_2 : 1 = female, 0 = male) on the risk of an anaphylactic allergic reaction during lumbar disk surgery. From a sample of $N = 44,913$ surgical procedures, the data in each category of the covariates are presented in Table 7.14.

7.3.1. Conduct a logistic regression analysis of the influence of route of administration on the odds of a reaction with no adjustment.

7.3.2. Then conduct a logistic regression analysis that is also adjusted for gender. Include the *covb* option in this analysis.

7.3.3. Fit an additional model that includes an interaction between route of administration and gender. The Wald or likelihood ratio test of the interaction effect is equivalent to the hypothesis of homogeneity among strata described in Section 4.6.

7.3.4. From the model in Problem 7.3.2, based on the estimated coefficients and the estimated covariance matrix of the estimates, compute the estimated probability of an allergic reaction and its 95% confidence limits from (7.17) for each of the following subpopulations:

1. Female with local anesthesia.
2. Male with local anesthesia.
3. Female with general anesthesia.
4. Male with general anesthesia.

Note that these probabilities may also be obtained from PROC LOGISTIC using the *ppred* option.

7.4 Use the *GLM* construction for binomial models described in Section 7.1.5 and the expressions (7.28)–(7.30) based on the chain rule. Then, for a two-covariate model as in Problem 7.1, derive the likelihood, log likelihood, the score estimating equations, the elements of the Hessian, and the observed and expected information matrices for:

7.4.1. An exponential risk model using the log link: $\log(\pi_i) = \alpha + \mathbf{x}'_i \boldsymbol{\beta}$.

7.4.2. A compound exponential risk model using the complementary log-log link: $\log[-\log(\pi_i)] = \alpha + \mathbf{x}'_i \boldsymbol{\beta}$.

7.4.3. A probit regression model using the inverse probit link: $\Phi(\pi_i) = \alpha + \mathbf{x}'_i \boldsymbol{\beta}$, where $\Phi(\cdot)$ is the cumulative normal distribution function, the inverse function for which is the probit, or $\pi_i = \Phi^{-1}(\alpha + \mathbf{x}'_i \boldsymbol{\beta})$.

7.4.4. Use PROC GENMOD to conduct an analysis of the data from Problem 4.9 using each of the above links.

7.5 Consider a model of the form $\log(y) = \alpha + \log(x)\beta$, where the coefficient equals the change in $\log(Y)$ per unit change in $\log(X)$.

7.5.1. Show that a c -fold difference in X , such as $x_1 = cx_0$, is associated with a c^β -fold change in Y . Thus, the percent change in Y per c -fold change in X is $100(c^\beta - 1)$ for $c > 0$.

7.5.2. Suppose that a model fit using covariate $\log(X)$ yields $\hat{\beta} = -2.1413$ with $S.E.(\hat{\beta}) = 0.9618$. Show that this implies that there is a 25.3% increase in the odds per 10% reduction in X , with 95% C.I. (2.74, 52.8%).

7.6 Consider a regression model with two binary covariates (X_1, X_2) and a structural relationship as expressed in (7.45) for some smooth link function $g(\cdot)$ with independent errors having mean zero. Then for the bivariate model we assume that $E(y | x_1, x_2) = g^{-1}(\alpha + x_1\beta_1 + x_2\beta_2)$. For simplicity we assume that X_1 and

X_2 are statistically independent and that X_2 has probabilities $P(x_2 = 1) = \phi$ and $P(x_2 = 0) = 1 - \phi$. Now consider the values of the coefficients when X_2 is dropped from the model to yield $E(y|x_1) = g^{-1}(\tilde{\alpha} + x_1\tilde{\beta}_1)$. Note that this notation is simpler than that used in Section 7.2.2.

7.6.1. Show that the coefficients in the reduced model can be obtained as

$$\begin{aligned} E(y|x_1) &= g^{-1}(\tilde{\alpha} + x_1\tilde{\beta}_1) = E_{x_2}[E(y|x_1, x_2)] \\ &= (1 - \phi)g^{-1}(\alpha + x_1\beta_1) + \phi g^{-1}(\alpha + x_1\beta_1 + \beta_2). \end{aligned} \quad (7.159)$$

7.6.2. Let $g(\cdot)$ be the identity function. Show that $\tilde{\alpha} = \alpha + \phi\beta_2$ and $\tilde{\beta}_1 = \beta_1$.

7.6.3. Let $g(\cdot)$ be the log function, $g^{-1}(\cdot) = \exp(\cdot)$. Show that $\tilde{\alpha} = \alpha + \log(1 - \phi + \phi e^{\beta_2})$ and $\tilde{\beta}_1 = \beta_1$.

7.6.4. Let $g(\cdot)$ be the logit function and $g^{-1}(\cdot)$ the inverse logit. Assume that X_1 is a binary variable. Show that

$$\begin{aligned} \tilde{\alpha} &= \log \frac{\pi_0}{1 - \pi_0} \\ \pi_0 &= E(y|x_1 = 0) = \frac{1 - \phi}{1 + e^{-\alpha}} + \frac{\phi}{1 + e^{-(\alpha + \beta_2)}} \end{aligned} \quad (7.160)$$

$$\begin{aligned} \tilde{\beta}_1 &= \log \frac{\pi_1}{1 - \pi_1} - \tilde{\alpha} \\ \pi_1 &= E(y|x_1 = 1) = \frac{1 - \phi}{1 + e^{-(\alpha + \beta_1)}} + \frac{\phi}{1 + e^{-(\alpha + \beta_1 + \beta_2)}}, \end{aligned} \quad (7.161)$$

where $\tilde{\beta}_1$, in general, does not equal β_1 . Note that $e^a/(1 + e^a) = (1 + e^{-a})^{-1}$.

7.6.5. Assume a logit model with $\alpha = 0.2$, $\beta_1 = 0.5$, and $\beta_2 = 1.0$, where $\phi = 0.4$. If X_2 is dropped from the model, show that $\tilde{\alpha} = 0.5637$ and $\tilde{\beta}_1 = 0.4777$.

7.6.6. Let $g(\cdot)$ be the complementary log-log function with the inverse link function $g^{-1}(\cdot) = \exp[-\exp(\cdot)]$. Show that

$$\begin{aligned} \tilde{\alpha} &= \log[-\log(\pi_0)] \\ \pi_0 &= E(y|x_1 = 0) = e^{-e^\alpha} \left[1 - \phi + \phi e^{-e^\alpha e^{\beta_2} - 1} \right] \end{aligned} \quad (7.162)$$

and that

$$\begin{aligned} \tilde{\beta}_1 &= \log[-\log(\pi_0)] - \tilde{\alpha} \\ \pi_1 &= E(y|x_1 = 1) = e^{-e^{\alpha + \beta_1}} \left[1 - \phi + \phi e^{-e^{\alpha + \beta_1} e^{\beta_2} - 1} \right], \end{aligned} \quad (7.163)$$

where $\tilde{\beta}_1$, in general, does not equal β_1 .

7.7 For the logistic regression model, show that under $H_0: \beta = 0$:

7.7.1. $U[\hat{\alpha}, \beta]_{|\beta=0}^T$ is as presented in (7.57).

- 7.7.2.** $I(\hat{\alpha}, \beta)_{|\beta=0}$ is as presented in (7.59).
- 7.7.3.** The efficient score test for H_0 is as presented in (7.60).
- 7.7.4.** Verify the computation of the score vector in Example 7.6 under H_0 , the corresponding information matrix and the score test value.
- 7.8** Now consider the log-risk model with a log link as in Problem 7.4.1. Generalizing the results of Problem 6.4, for a model with p covariates \mathbf{x} , intercept α and coefficient vector β , derive the expression for the model score test under $H_0: \beta = 0$.

7.9 Direct Adjustment (Gastwirth and Greenhouse, 1995). In model-based direct adjustment, the model obtained from one sample is applied to a different sample of subjects. For example, a logistic model might be obtained from a sample of M males and then applied to a sample of N females to determine whether the number of responses among females differs from what would be expected if the male risk function also applied to females. In the sample of females, the observed number of responses is $O = \sum_{i=1}^N y_i$. Then the expected number of responses is estimated as $\hat{E} = \sum_{i=1}^N \hat{\pi}_i$, where $\hat{\pi}_i$ is provided by the model derived from the sample of males. Since the model coefficients are obtained from an independent sample, then

$$V(O - \hat{E}) = V\left(\sum_{i=1}^N y_i\right) + V\left(\sum_{i=1}^N \hat{\pi}_i\right) = V(O) + V(\hat{E}), \quad (7.164)$$

which requires $V(\hat{\pi}_i)$. Now do the following:

- 7.9.1.** Show that $V(O) = \sum_i \pi_i(1 - \pi_i)$, where $\pi_i = E(y_i)$.
- 7.9.2.** Under the hypothesis that the male model also applies to females, show that the male model provides an estimate of $V(O)$ as

$$\hat{V}(O) = \sum_{i=1}^N \hat{\pi}_i(1 - \hat{\pi}_i). \quad (7.165)$$

- 7.9.3.** Then show that

$$V\left[\hat{E}\right] = \sum_i V(\hat{\pi}_i). \quad (7.166)$$

- 7.9.4.** The estimated probabilities are based on the linear predictor for the i th subject, $\hat{\eta}_i = \tilde{\mathbf{x}}_i' \boldsymbol{\theta} = \alpha + \sum_{j=1}^p x_{ij} \hat{\beta}_j$ in the notation of (7.15). Use the δ -method to show that

$$\begin{aligned} \hat{V}(\hat{\pi}_i) &\cong \left(\frac{d\hat{\pi}_i}{d\hat{\eta}_i}\right)^2 \hat{V}(\hat{\eta}_i) = \frac{(e^{-\hat{\eta}_i})^2}{(1 + e^{-\hat{\eta}_i})^4} \tilde{\mathbf{x}}_i' \hat{\Sigma}_{\boldsymbol{\theta}} \tilde{\mathbf{x}}_i \\ &= \hat{\pi}_i^2 (1 - \hat{\pi}_i)^2 \tilde{\mathbf{x}}_i' \hat{\Sigma}_{\boldsymbol{\theta}} \tilde{\mathbf{x}}_i. \end{aligned} \quad (7.167)$$

- 7.9.5.** Assume that a new anesthetic drug has been developed and that a new study is conducted that shows the following numbers of allergic reactions with local and general anesthesia among males and females in the same format as presented for the drug chymopapain in Table 7.14:

Stratum j	# Allergic Reactions a_j	# Surgeries n_j	(Intercept, Route, Gender) (1, x_{1j} , x_{2j})
1	53	18,497	(1, 0, 0)
2	85	11,286	(1, 0, 1)
3	19	7,404	(1, 1, 0)
4	34	5,369	(1, 1, 1)

Apply the logistic regression model obtained from the study of chymopapain obtained in Problem 7.3, with its estimated covariance matrix, to these data obtained for the new anesthetic drug. Estimate $O - E$ and its variance under the null hypothesis that the risk of allergic reactions with chymopapain also applies to the new agent. Compute a Wald test of this hypothesis.

7.10 Use the DCCT nephropathy data of Table 7.5 that can be obtained online (see the Preface).

7.10.1. Verify the computation of the matrices $\widehat{J}(\widehat{\theta})$ and $\widehat{\Sigma}_R(\widehat{\theta})$ presented in Example 7.8, the robust confidence limits, and Wald tests for each of the model parameters.

7.10.2. Verify the computation of the matrices $\widehat{J}(\widehat{\theta}_0)$ and $\widehat{\Sigma}_R(\widehat{\theta}_0)$ and the computation of the robust efficient score test of the model.

7.11 Consider the determination of sample size and power for a logistic regression model using the study design in Problem 4.11. Although the odds ratios may differ among strata, we desire the sample size for a stratified-adjusted analysis with an overall treatment group odds ratio. Assume that all tests are conducted at the 0.05 level.

7.11.1. Determine the corresponding logistic regression model parameters by generating a large data set with cell frequencies proportional to the projected sample fractions and with numbers of positive responses proportional to the response probabilities within each cell.

7.11.2. Then compute the covariance matrix Ω of the Bernoulli variables and $V(\widehat{\theta})$ as in (7.77).

7.11.3. Now, consider a Wald test of the adjusted group effect on 1 df based on a contrast vector $C = (0 \ 0 \ 0 \ 1)^T$. Compute the noncentrality factor K^2 as in (7.79).

7.11.4. Determine the sample size required to provide 90% power for this 1 df Wald test.

7.11.5. Likewise, consider the Wald model test on 3 df of the null hypothesis $H_0: \beta = 0$. Here the matrix C is a 4×3 matrix of rank 3 where $C = (0 \ 0 \ 0 // 1 \ 0 \ 0 // 0 \ 1 \ 0 // 0 \ 0 \ 1)$ and “//” denotes concatenation of row vectors. Compute the noncentrality factor K^2 as in (7.79).

7.11.6. Determine the sample size required to provide 90% power for this 3 df Wald model test.

7.11.7. Suppose that a sample size of $N = 220$ is proposed. What level of power does this provide for the 1 df test of treatment group?

- 7.11.8.** What level of power does this provide for the 3 *df* model test?
- 7.12** Consider the case-control study of ischemic heart disease that is presented in Example 7.5. The effect of hypertension, adjusting for smoking and hyperlipidemia, is an odds ratio of 1.386 that is not statistically significant at the 0.05 level based on the Wald test (i.e., on the 95% confidence intervals).
- 7.12.1.** Treating the sample frequencies within each cell as fixed, and treating the intercept and coefficients for smoking and hyperlipidemia as fixed (equal to the estimates shown in the example), what level of power was there to detect an odds ratio of 1.5 for those with versus without hypertension using the 1 *df* Wald test?
- 7.12.2.** What sample size would be required to provide 90% power to detect an odds ratio for hypertension of 2.0?
- 7.13** Consider models with an interaction between two covariates as in Section 7.4.
- 7.13.1.** From the interaction model in (7.88) for the ulcer clinical trial data in Example 4.1, compute the odds and odds ratios for the cells of the table in (7.89) for drug versus placebo within each stratum.
- 7.13.2.** Fit a similar stratum \times group interaction model to:
1. The religion and mortality data of Example 4.6.
 2. The pre-eclampsia data of Example 4.24.
 3. The smoking and byssinosis data of Problem 4.9.
 4. The allergic reactions data of Table 7.14, treating route of administration as the exposure factor and gender as the stratification factor.
- 7.13.3.** For each, compute the odds and odds ratios for the principal exposure or treatment group versus the comparator within each stratum.
- 7.13.4.** For each, compute the Wald and likelihood ratio tests of homogeneity among strata.
- 7.13.5.** Using the DCCT interaction model presented in Example 7.12, compute the odds ratio for intensive versus conventional treatment over the range 3–14 years duration of diabetes that corresponds approximately to the 5th to 95th percentiles of the distribution of duration. Display the estimates in a figure such as Figure 7.1.
- 7.13.6.** For this model, also compute the odds ratio per unit increase in blood pressure over the range of HbA_{1c} values of 7–12 %, also corresponding approximately to the 5th to 95th percentiles. Display the estimates in a figure.
- 7.13.7.** Consider a model with two covariates (X_1, X_2) where $X_1 = \log(Z)$ and there is an interaction between X_1 and X_2 . Show that the odds ratio per *c*-fold change in Z given the value of X_2 is obtained as
- $$\widehat{OR} = e^{\widehat{\beta}_1 + x_2 \widehat{\beta}_{12}}. \quad (7.168)$$
- 7.13.8.** Using the DCCT data, fit an interaction model as in Example 7.12 using $X_2 = \log(\text{HbA}_{1c})$.
- 7.13.9.** Conduct a 2 *df* test of the effect of HbA_{1c} .
- 7.13.10.** Compute the odds ratio per 10% reduction in HbA_{1c} , and its 95% confidence limits, over the range of blood pressure values and display the estimates in a figure.

7.14 For a logistic model, using expressions for the measures of explained variation, do the following:

7.14.1. Show that the model fits perfectly when $\hat{\pi}_i y_i + (1 - \hat{\pi}_i)(1 - y_i) = 1$ for all $i = 1, \dots, N$. In this case, since $\hat{\pi} = m_1/N$, also show that $\hat{\rho}_{risk}^2 = 1$.

7.14.2. Conversely, show that the model explains none of the variation in risk when $\hat{\pi}_i = \hat{\pi}$ for all subjects, irrespective of the covariate values, and that in this case $\hat{\rho}_{risk}^2 = 0$.

7.14.3. Noting that $\sum_i y_i = \sum_i \hat{\pi}_i$, show that the following equality applies:

$$\sum_i y_i \log(\hat{\pi}_i) + (1 - y_i) \log(1 - \hat{\pi}_i) = \sum_i \hat{\pi}_i \log(\hat{\pi}_i) + (1 - \hat{\pi}_i) \log(1 - \hat{\pi}_i).$$

7.14.4. From (A.202), show that the fraction of explained residual variation with squared error loss in a logistic regression model is

$$R_{e^2, resid}^2 = \frac{\sum_i \left(\hat{\pi}_i - \hat{\pi} \right)^2}{\sum_i \left(y_i - \hat{\pi} \right)^2}, \quad (7.169)$$

where $\hat{y}_i = \hat{\pi}_i$ and $\bar{y} = \bar{\pi}$.

7.14.5. Noting that $\hat{\pi} = \sum_i y_i/N$, show that $R_{resid}^2 = \hat{\rho}_{risk}^2$ in (7.95).

7.14.6. Show that $R_{E, resid}^2$ in (7.106) equals $\hat{\rho}_E^2$ in (7.105).

7.14.7. Show that $R_{E, resid}^2$ in (7.106) equals R_E^2 in (7.107).

7.14.8. For a single 2×2 table with a single binary covariate for treatment group, based on Problem 6.2.13, show that the entropy R^2 equals

$$R_E^2 = \frac{\log \left[\frac{a^a b^b c^c d^d N^N}{n_1^{n_1} n_2^{n_2} m_1^{m_1} m_2^{m_2}} \right]}{\log \left[\frac{N^N}{m_1^{m_1} m_2^{m_2}} \right]}. \quad (7.170)$$

7.14.9. Among the many measures of association in a 2×2 table, Goodman and Kruskal (1972) suggested the uncertainty coefficient that is based on the ability to predict column membership (the response) from knowledge of the row membership (the exposure group). For the 2×2 table, this coefficient can be expressed as

$$U(C|R) = \frac{\sum_{i=1}^2 n_i \log \frac{n_i}{N} + \sum_{j=1}^2 m_j \log \frac{m_j}{N} - \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \log \frac{n_{ij}}{N}}{\sum_{j=1}^2 m_j \log \frac{m_j}{N}}, \quad (7.171)$$

where the $\{n_{ij}\}$, $\{n_i\}$, and $\{m_j\}$ are the cell frequencies, row (group) totals, and column (response) totals, respectively. Show that $R_E^2 = U(C|R)$.

7.15 Collett (1991) presents the data given in Table 7.15 from a study of factors associated with local metastases to the lymph nodes (nodal involvement) in subjects

Table 7.15 SAS program to input prostate cancer data from Collett (1991), reproduced with permission.

```

data prostate; input case age acid xray size grade nodalinv @@;
lacd=log(acid);
datalines;
1 66 .48 0 0 0 0 2 68 .56 0 0 0 0 3 66 .50 0 0 0 0
4 56 .52 0 0 0 0 5 58 .50 0 0 0 0 6 60 .49 0 0 0 0
7 65 .46 1 0 0 0 8 60 .62 1 0 0 0 9 50 .56 0 0 1 1
10 49 .55 1 0 0 0 11 61 .62 0 0 0 0 12 58 .71 0 0 0 0
13 51 .65 0 0 0 0 14 67 .67 1 0 1 1 15 67 .47 0 0 1 0
16 51 .49 0 0 0 0 17 56 .50 0 0 1 0 18 60 .78 0 0 0 0
19 52 .83 0 0 0 0 20 56 .98 0 0 0 0 21 67 .52 0 0 0 0
22 63 .75 0 0 0 0 23 59 .99 0 0 1 1 24 64 1.87 0 0 0 0
25 61 1.36 1 0 0 1 26 56 .82 0 0 0 1 27 64 .40 0 1 1 0
28 61 .50 0 1 0 0 29 64 .50 0 1 1 0 30 63 .40 0 1 0 0
31 52 .55 0 1 1 0 32 66 .59 0 1 1 0 33 58 .48 1 1 0 1
34 57 .51 1 1 1 1 35 65 .49 0 1 0 1 36 65 .48 0 1 1 0
37 59 .63 1 1 1 0 38 61 1.02 0 1 0 0 39 53 .76 0 1 0 0
40 67 .95 0 1 0 0 41 53 .66 0 1 1 0 42 65 .84 1 1 1 1
43 50 .81 1 1 1 1 44 60 .76 1 1 1 1 45 45 .70 0 1 1 1
46 56 .78 1 1 1 1 47 46 .70 0 1 0 1 48 67 .67 0 1 0 1
49 63 .82 0 1 0 1 50 57 .67 0 1 1 1 51 51 .72 1 1 0 1
52 64 .89 1 1 0 1 53 68 1.26 1 1 1 1
;

```

with prostate cancer, indicated by the variable *nodalinv* (= 1 if yes, 0 if no). The risk factors considered are age at diagnosis (*age*), the level of acid phosphatase in serum in King-Armstrong units (*acid*) that is a measure of the degree of tissue damage, positive versus negative signs of nodal involvement on x-ray (*xray* = 1 if positive, 0 if not), size of the tumor based on rectal examination (*size* = 1 if large, 0 if not), and the histologic grade of severity of the lesion based on a tissue biopsy (*grade* = 1 if serious, 0 if not). The variable *acid* is log transformed (*lacd*). This data set is also included in the SAS (1995) text on logistic regression and is available online (see the Preface). Conduct the following analyses of these data:

7.15.1. Use PROC FREQ of *nodalinv**(*xray size grade*) to obtain an unadjusted estimate of the odds ratio and the associated logit confidence limits associated with each of the binary covariates.

7.15.2. Conduct a logistic regression analysis of the odds of nodal involvement using PROC LOGISTIC with *age*, *lacd*, *xray*, *size*, and *grade* as covariates. Interpret all coefficients in terms of the associated odds ratio with their asymmetric confidence limits.

7.15.3. For *age*, also compute $\Delta \log(\text{odds})$ per five years lower age and the associated confidence limits.

7.15.4. Since *lacd* is $\log(\text{acid})$, describe the effect of acid phosphatase on the risk of nodal involvement in terms of the percent change in odds per 10% larger acid phosphatase, with the associated 95% confidence limits (see Problem 7.5.1).

7.15.5. Compare the unadjusted and adjusted odds ratio for *xray*, *size*, and *grade*.

7.15.6. Conduct a logistic regression analysis using PROC GENMOD to compute likelihood ratio tests for each covariate in the model. Compare these to the Wald tests.

7.15.7. Compute R^2_{resid} , R^2_E , and R^2_{LR} for these data. To obtain the corrected sum of squares of the predicted probabilities, use the following statements after the PROC LOGISTIC *model* statement:

```
output out=two pred=pred;
proc univariate data=two; var pred;
```

7.15.8. Use PROC GENMOD to compute the robust information sandwich estimator of the covariance matrix of the estimates and the associated robust 95% confidence limits and the robust Wald tests for each covariate (see Section 8.4.2). Compare these to those obtained under the assumed model.

7.15.9. Write a program to verify the computation of the robust information sandwich estimate of the covariance matrix of the estimates.

7.15.10. Write a program to evaluate the score vector $\mathbf{U}(\hat{\theta}_0)$ under the model null hypothesis $H_0: \beta = 0$, and to compute the robust information sandwich estimate of the covariance matrix under this null hypothesis. Use these to compute the robust model score test that equals $X^2 = 15.151$ on 5 *df* with $p \leq 0.0098$. Compare this to the likelihood ratio and score tests obtained under the model.

7.15.11. Fit a model with pairwise interactions between *age* and *lacd*, and between *size* and *grade*. Even though the interaction between *age* and *lacd* is not significant at the 0.05 level, provide an interpretation of both interaction effects as follows.

1. Compute the odds ratio for a serious versus not serious grade of cancer as a function of the size category of the tumor. Compute the variance of each log odds ratio and then the asymmetric 95% confidence limits for each odds ratio.

2. Compute the odds ratio for a large versus not tumor size as a function of the grade category or the tumor. Compute the variance of each log odds ratio and then the asymmetric 95% confidence limits for each odds ratio.

3. Compute the odds ratio associated with a 10% higher level of the acid phosphatase as a function of the age of the patient, using at least ages 50, 55, 60, and 65, and for each compute the 95% asymmetric confidence limits.

4. Compute the odds ratio associated with a five-year higher age as a function of the level of log acid phosphatase, using at least the values of acid phosphatase of 0.5, 0.65, 0.8, and 1.0, and for each compute the 95% asymmetric confidence limits.

7.15.12. Now fit a model with interactions between *lacd* and *grade*, and between *size* and *grade*.

1. Compute the odds ratio for serious versus not serious grade of cancer as a function of the size category of the tumor and the level of log acid phosphatase, using at least the above specified values.
2. Compute the variance of each log odds ratio and then the asymmetric 95% confidence limits for each odds ratio.
3. Use a model with the *test* option to compute a multiple degree-of-freedom test of the overall effect of *lacd*, *grade*, and *size*.

7.16 Consider the logistic regression model for matched sets in Section 7.6.

- 7.16.1. Derive the likelihood $L(\alpha, \beta)$ in (7.113).
- 7.16.2. Then conditioning on the m_{1i} responses with respective covariate values as in (7.115), derive the conditional logistic likelihood for matched sets in (7.116).
- 7.16.3. Show that the likelihood reduces to (7.117).
- 7.16.4. Derive the expression for the score equation for β_k presented in (7.121).
- 7.16.5. Derive the expressions for the elements of the information matrix presented in (7.122) and (7.123).

7.17 For the matched retrospective study described in Section 7.6.3, show that the retrospective conditional likelihood in (7.125) equals the prospective conditional likelihood in (7.116).

7.18 Breslow and Day (1980) present data from a matched case-control study with four controls per case from a study of the risk of endometrial cancer among women who resided in a retirement community in Los Angeles (Mack et al., 1976). These data are available online (see the Preface, reproduced with permission). Each of the 63 cases was matched to four controls within one year of age of the case who had the same marital status and entered the community at about the same time. The variables in the data set are

caseset: the matched set number.

case: 1 if case, 0 if control, in sets of 5 (1 + 4).

age: in years (a matching variable).

gbdx: history of gallbladder disease (1 = yes, 0 = no).

hypert: history of hypertension, (1 = yes, 0 = no).

obese: (1 = yes, 0 = no, 9 = unknown). Breslow and Day combine the no and unknown categories in their analyses.

estrogen: history of estrogen use, (1 = yes, 0 = no).

dose: dose of conjugated estrogen use: 0 = none, 1 = 0.3; 2 = (0.301-0.624); 3 = 0.625; 4 = (0.626-1.249); 5 = 1.25; 6 = (1.26-2.50); 9 = unknown.

dur: duration of conjugated estrogen use in months, values > 96 are truncated at 96, 99 = unknown.

nonest: history of use of nonestrogen drug (1 = yes, 0 = no).

Note that conjugated estrogen use is a subset of all estrogen use. Also the variables for dose and duration are defined only for *conjugated* estrogen use. Thus, for some analyses below an additional variable *conjest* must be defined as 0 for *dose* = 0, 1 otherwise. Those subjects with an unknown dose are assumed to have used conjugated estrogens.

Also, in the Breslow and Day (1980) analyses of these data, the *dose* categories 3 to 6 were combined, and those with unknown dose treated as missing. This yields a variable with four categories: none (0), low (1), middle (2), and high (3).

From these data, perform the following analyses: In each case, state the model and describe the interpretation of the regression coefficients.

7.18.1. Fit the conditional logistic model using PROC PHREG to assess the effect of estrogen use (yes vs. no) on the odds of having cancer with no other covariates (unadjusted) and then adjusted for other covariates (*gbdx*, *hypert*, *obese*, *nonest*).

7.18.2. Assess the effect of the dose of conjugated estrogen on risk unadjusted and then adjusted for the other four covariates. Do so using the four groupings of dose described above (*none*, *low*, *middle*, *high*). Since this is a categorical variable, use a 3 *df* test and three odds ratios for the *low*, *middle*, and *high* dose categories versus none. Do so using two different models, one using three indicator variables with *none* as the reference, another using a *class* statement with *none* as the reference.

7.18.3. The analyses in Problem 7.18.2 compare each dose level to the nonusers. Now perform an analysis comparing the high and middle doses to the low dose. This can be done with a new class variable where the reference category is the low dose, or using three binary variables for each of the other categories versus the low dose (including the nonusers).

7.18.4. Now consider the effect of the duration of conjugated estrogen use including both *conjest* and *dur* in the model because we wish to describe the duration effect only among the estrogen users. Fit the model with both covariates unadjusted and then adjusted for other covariates.

7.18.5. Using the model with just *conjest* and *dur*, show the effects of both variables simultaneously by deriving the expressions for the logit of the probability for a nonuser, and then for a user with increasing durations of conjugated estrogen use (e.g., 20, 40, 60, and 80 months of such use). Compute the estimated logits and corresponding probabilities. Note that "duration of drug use" is actually a mixture of two variables: the binary variable for any drug use (yes or no), and if yes, the duration of such use.

Analysis of Count Data

Thus far we have principally considered the risk of a binary outcome (Y), such as a positive or negative response ($y = 1$ or 0) over an assumed fixed period of exposure for each subject. In many settings, however, multiple events may occur in which case Y represents the number of events per unit of exposure, where $y = 0, 1, 2, \dots$. Although the exposure period may vary from subject to subject, the frequency of events is usually summarized as a rate or number per year of exposure, or per 100 patient years, and so on.

Common examples include the rate of serious adverse events per year among subjects in a pharmaceutical clinical trial, seizures per year among subjects with epilepsy; infections per year among schoolchildren; hospitalizations per year among patients with cardiovascular disease; and episodes of hypoglycemia (low blood sugar) per year among patients with diabetes. Although time is the unit of exposure in these examples, it need not always be the case. Other examples are the rate of infections per hospitalization, cavities per set of teeth, and defects or abnormalities per item.

This chapter describes methods for the analysis of such rate data analogous to those presented in previous chapters for prevalence and incidence data. Since the majority of these applications occur in the setting of a prospective study, we omit discussion of methods for the analysis of retrospective studies. The presentation is also limited to a binary outcome and does not consider polychotomous outcomes.

8.1 EVENT RATES AND THE HOMOGENEOUS POISSON MODEL

8.1.1 Poisson Process

To simplify notation, let $d_j = y_j$ denote the observed number of events during t_j units of exposure (months, years, etc.) for the j th in a sample of N subjects ($j = 1, \dots, N$), and let $d_\bullet = \sum_{j=1}^N d_j$ and $t_\bullet = \sum_{j=1}^N t_j$ denote the respective sums. Although the actual event times may also be observed, here we focus only on the number of events, not their times. We then assume that the sequence of events and the aggregate count for each subject is generated by a *Poisson process*. The Poisson process is described in any general reference on stochastic processes, such as Cox and Miller (1965), Karlin and Taylor (1975), and Ross (1983), among many. The Poisson process specifies that the underlying risk or rate of events over time is characterized by the *intensity* $\alpha(t)$, possibly varying in t . The intensity of the process is the instantaneous probability of the event among those at risk at time t . Now, let $N(t)$ denote the *counting process*, which is the cumulative number of events observed over the interval $(0, t]$. Then one of the features of the Poisson process with intensity $\alpha(t)$ is that probability of any number of events, say $d = m$, in an interval $(t, t + s]$ is given by the Poisson probability

$$P\{[N(t + s) - N(t)] = m\} = \frac{e^{-\Lambda_{(t,t+s)}} (\Lambda_{(t,t+s)})^m}{m!}, \quad t > 0, s > 0 \quad (8.1)$$

with rate parameter

$$\Lambda_{(t,t+s)} = \int_t^{t+s} \alpha(u) \, du \quad (8.2)$$

equal to the *cumulative intensity* over the interval. This cumulative intensity or rate parameter can also be expressed as $\Lambda_{(t,t+s)} = \lambda_{(t,t+s)}s$, where the *mean or linearized rate* is obtained by the mean value theorem as

$$\lambda_{(t,t+s)} = \frac{\Lambda_{(t,t+s)}}{s}. \quad (8.3)$$

Therefore, the probability of events over any interval of time can be obtained from the Poisson distribution with rate parameter $\theta = \Lambda_{(t,t+s)} = \lambda_{(t,t+s)}s$.

8.1.2 Doubly Homogeneous Poisson Model

In the most general case, the Poisson process may be time heterogeneous where the intensity $\alpha(t)$ varies over time, also called a nonhomogeneous Poisson process. In the simplest case, the intensity of the process may be a fixed constant $\alpha(t) = \lambda$ for $\forall t$, in which case the process is homogeneous over time. In this *homogeneous Poisson process* the number of events in an interval of length t is distributed as *Poisson*(λt) or

$$\Pr(d; t, \lambda) = P_o(d; t, \lambda) = \frac{e^{-\lambda t} (\lambda t)^d}{d!}, \quad d \geq 0, t > 0. \quad (8.4)$$

Herein we also make the additional assumption that this constant intensity applies to all subjects in the population, that we call the *doubly homogeneous Poisson model*. Under this model, the likelihood function for a sample of N independent observations is

$$L(\lambda) = \prod_{j=1}^N \frac{e^{-\lambda t_j} (\lambda t_j)^{d_j}}{d_j!} \quad (8.5)$$

given the set of exposure times $\{t_1, \dots, t_N\}$, that is, conditioning on the exposure times as fixed constants. The log likelihood is

$$\ell(\lambda) = \sum_{j=1}^N [(-\lambda t_j) + d_j \log(t_j) + d_j \log(\lambda) - \log(d_j!)] \quad (8.6)$$

Then the efficient score is

$$U(\lambda) = \frac{d\ell}{d\lambda} = \frac{1}{\lambda} \sum_j [d_j - \lambda t_j] = \frac{1}{\lambda} \sum_j [d_j - E(d_j|\lambda, t_j)], \quad (8.7)$$

where d_j and $E(d_j|\lambda, t_j)$ are the observed and expected number of events, respectively. The maximum likelihood estimating equation $U(\lambda) = 0$ then implies that $\sum_j d_j = \sum_j \lambda t_j$, so that the *MLE* is

$$\hat{\lambda} = \frac{\sum_j d_j}{\sum_j t_j} = \frac{d_{\bullet}}{t_{\bullet}}, \quad (8.8)$$

which equals the total number of events (d_{\bullet}) divided by the total period of exposure (t_{\bullet}). This estimator is called the *crude rate*, the *linearized rate* estimate, or the *total time on test* estimate, the latter term arising in reliability theory.

This estimator can also be expressed as a weighted *mean rate*. Let the rate for the j th subject be denoted as $r_j = d_j/t_j$. Under the constant intensity assumption, $E(r_j) = \lambda$ for all subjects. Then it is readily shown that the *MLE* in (8.8) can be expressed as a weighted mean of the r_j , weighted by the t_j

$$\hat{\lambda} = \bar{r} = \frac{\sum_j t_j r_j}{\sum_j t_j} = \frac{d_{\bullet}}{t_{\bullet}}. \quad (8.9)$$

The variance of the estimate can be easily derived as follows. Under the Poisson assumption $E(d_j) = V(d_j) = \lambda t_j$, the expression for the *MLE* then implies that

$$\begin{aligned} V(\hat{\lambda}) &= \left(\frac{1}{\sum_j t_j} \right)^2 \sum_j V(d_j) = \frac{\lambda}{\sum_j t_j} = \frac{\lambda}{t_{\bullet}} \\ &= \frac{\lambda^2}{\sum_j \lambda t_j} = \frac{\lambda^2}{\sum_j E(d_j)} = \frac{\lambda^2}{E(d_{\bullet})}. \end{aligned} \quad (8.10)$$

The latter expression shows that the variance of the estimate is λ^2 divided by the expected number of events. Thus, for given λ the variance is inversely proportional to the total exposure time and also to the total expected number of events.

This expression for the variance can also be derived as the inverse of Fisher's information. The observed information is

$$i(\lambda) = \frac{-dU(\lambda)}{d\lambda} = \frac{\sum_j d_j}{\lambda^2} = \frac{d_{\bullet}}{\lambda^2}, \quad (8.11)$$

and the expected information $I(\lambda) = E[i(\lambda)]$ is

$$I(\lambda) = E\left[\frac{\sum_j d_j}{\lambda^2}\right] = \frac{E(d_{\bullet})}{\lambda^2} = \frac{t_{\bullet}}{\lambda}. \quad (8.12)$$

Thus, the total information in the data is proportional to the total expected number of events $E(d_{\bullet})$ conditional on the $\{t_j\}$. Then the variance of the estimate is as shown in (8.10) and is consistently estimated as

$$\widehat{V}(\widehat{\lambda}) = \frac{\widehat{\lambda}^2}{d_{\bullet}} = \frac{\widehat{\lambda}}{t_{\bullet}}. \quad (8.13)$$

Asymptotically, since $\widehat{\lambda}$ is the *MLE*, it follows that

$$(\widehat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}[0, V(\widehat{\lambda})]. \quad (8.14)$$

This result can also be derived from the central limit theorem conditioning on the $\{t_j\}$ since $\widehat{\lambda}$ is the weighted mean rate. Although this provides large sample tests of significance and confidence limits, since λ is positive a more accurate approximation is provided by $\widehat{\eta} = \log(\widehat{\lambda})$. From the δ -method it is readily shown that

$$V(\widehat{\eta}) = \left(\frac{1}{\lambda}\right)^2 V(\widehat{\lambda}) = \frac{1}{\sum_j E(d_j)} = \frac{1}{E(d_{\bullet})}, \quad (8.15)$$

so that the estimated variance is $\widehat{V}(\widehat{\eta}) = d_{\bullet}^{-1}$. This then yields asymmetric confidence limits for the assumed constant intensity.

8.1.3 Relative Risks

Now assume that we have independent samples of n_1 and n_2 subjects drawn at random from two separate populations, such as exposed versus not exposed, or treated versus control. We then wish to assess the relative risk of events between the two groups, including possible multiple or recurrent events. Under a doubly homogeneous Poisson model, this relative risk is described as simply the ratio of the assumed constant intensities

$$RR = \lambda_1/\lambda_2 \cong \widehat{\lambda}_1/\widehat{\lambda}_2, \quad (8.16)$$

where $\widehat{\lambda}_i = d_{i\bullet}/t_{i\bullet}$ is the estimated rate in the i th group of n_i subjects with $d_{i\bullet} = \sum_{j=1}^{n_i} d_{ij}$ and $t_{i\bullet} = \sum_{j=1}^{n_i} t_{ij}$, $i = 1, 2$. Thus, we can use $\log(\widehat{RR})$ as the basis for confidence intervals or use $\widehat{\lambda}_1 - \widehat{\lambda}_2$ for a statistical test.

Let $\eta_i = \log(\lambda_i)$ and $\theta = \log(RR) = \eta_1 - \eta_2$. Then

$$\hat{\theta} = \log(\widehat{RR}) = \hat{\eta}_1 - \hat{\eta}_2, \quad (8.17)$$

where $\hat{\eta}_i = \log(\hat{\lambda}_i)$, $i = 1, 2$. The variance of the estimate is

$$V(\hat{\theta}) = V(\hat{\eta}_1) + V(\hat{\eta}_2) = \frac{1}{E(d_{1\bullet})} + \frac{1}{E(d_{2\bullet})}, \quad (8.18)$$

which is estimated as

$$\hat{V}(\hat{\theta}) = \frac{1}{d_{1\bullet}} + \frac{1}{d_{2\bullet}} = \frac{d_{1\bullet} + d_{2\bullet}}{d_{1\bullet}d_{2\bullet}}. \quad (8.19)$$

This provides large sample confidence limits for θ and asymmetric confidence limits for the relative risk.

An efficient large sample test of $H_0: \lambda_1 = \lambda_2 = \lambda$ is then obtained as

$$Z = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\hat{V}(\hat{\lambda}_1 - \hat{\lambda}_2 | H_0)}}, \quad (8.20)$$

using an estimate of the variance under the null hypothesis. This null variance is defined as

$$V(\hat{\lambda}_1 - \hat{\lambda}_2 | H_0) = V(\hat{\lambda}_1 | H_0) + V(\hat{\lambda}_2 | H_0). \quad (8.21)$$

From (8.10)

$$V(\hat{\lambda}_i | H_0) = \frac{\lambda^2}{E(d_{i\bullet})}_{|\lambda_i=\lambda} = \frac{\lambda}{t_{i\bullet}}. \quad (8.22)$$

Under H_0 the *MLE* of the assumed common rate is

$$\hat{\lambda} = \frac{d_{1\bullet} + d_{2\bullet}}{t_{1\bullet} + t_{2\bullet}}. \quad (8.23)$$

Therefore, the null variance is estimated as

$$\hat{V}(\hat{\lambda}_1 - \hat{\lambda}_2 | H_0) = \hat{\lambda} \left[\frac{1}{t_{1\bullet}} + \frac{1}{t_{2\bullet}} \right] = \frac{d_{1\bullet} + d_{2\bullet}}{t_{1\bullet}t_{2\bullet}}. \quad (8.24)$$

The resulting large sample test statistic is

$$Z = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{(d_{1\bullet} + d_{2\bullet})/(t_{1\bullet}t_{2\bullet})}}, \quad (8.25)$$

which is asymptotically distributed as standard normal under H_0 . In a problem it is also shown that this test is the efficient score test for the effect of a single binary covariate for treatment group in a Poisson regression model.

An asymptotically equivalent large sample test can be obtained as a test of H_0 : $\theta = \log(RR) = 0$ using

$$Z = \frac{\hat{\theta}}{\sqrt{\hat{V}(\hat{\theta}|H_0)}}, \quad (8.26)$$

where the null variance is estimated as

$$\hat{V}(\hat{\theta}|H_0) = \frac{1}{\hat{\lambda}} \left[\frac{1}{t_{1\bullet}} + \frac{1}{t_{2\bullet}} \right]. \quad (8.27)$$

Using a Taylor's expansion about λ , it is readily shown that the two tests are asymptotically equivalent.

Example 8.1 *Hypoglycemia in the DCCT*

In the Diabetes Control and Complications Trial (DCCT, see Section 1.5), the major potential adverse effect of intensive diabetes therapy with the technology then available was an episode of hypoglycemia where the blood glucose level falls too low, at which point the patient becomes dizzy, disoriented, and may pass out. Here we use the data from the secondary cohort of 715 patients with more advanced and longer duration diabetes. Table 8.1 presents a partial listing of the data that also includes covariate information (discussed later). For each subject this shows the treatment group assignment ($int: i = 1$ for intensive, 0 for conventional), the number of events ($d_{i\bullet}$), the years of exposure or follow-up ($t_{i\bullet}$), and the corresponding rate of events per year ($r_i = rate$).

The following is a summary of the overall incidence of severe hypoglycemia in the two treatment groups within this cohort:

	<i>Intensive</i> $i = 1$	<i>Conventional</i> $i = 2$	<i>Total</i>
n_i	363	352	715
Events ($d_{i\bullet}$)	1723	543	2266
Exposure time in years ($t_{i\bullet}$)	2598.5	2480.2	5078.7
$\hat{\lambda}_i$ per patient year (PY)	0.6631	0.2189	0.4462
$\hat{\lambda}_i$ per 100 PY	66.3	21.9	44.6

The crude rates are conveniently expressed as a number per 100 patient years. This is sometimes described as the percent of patients per year. Such a description is appropriate if there are no recurrent events, such as in survival analysis. Here, however, some patients experienced as many as 20 or more episodes of severe hypoglycemia, in which case it is inappropriate to describe these rates in terms of a percent of patients per year.

The ratio of these crude rates yields $\widehat{RR} = 3.029$ and $\widehat{\theta} = \log(\widehat{RR}) = 1.108$ with estimated variance $\hat{V}(\widehat{\theta}) = 1/1723 + 1/543 = 0.002422$ and $S.E.(\widehat{\theta}) = 0.0492$. This yields asymmetric 95% confidence limits on RR of (2.75, 3.335). Under the

Table 8.1 Episodes of severe hypoglycemia for a subset of subjects in the DCCT.

INT	D	T	RATE	INSULIN	DUR	FEM	ADU	BCV	HBAE	HXC
1	5	9.50582	0.52599	0.33557	110	0	1	0.01	9.74	0
0	9	9.47296	0.95007	0.71197	100	1	1	0.04	8.90	0
1	0	9.45106	0.00000	0.55487	119	1	1	0.01	9.99	0
1	6	9.24025	0.64933	0.58361	111	0	1	0.01	6.94	0
0	3	9.37714	0.31993	0.40984	99	1	1	0.01	10.94	0
0	1	9.37988	0.10661	0.48409	44	0	1	0.18	9.56	0
1	1	9.43190	0.10602	0.57143	180	0	1	0.01	9.67	0
0	0	8.09035	0.00000	0.34783	60	1	1	0.04	12.75	0
1	0	7.88775	0.00000	0.75472	112	0	1	0.04	8.57	0
1	3	7.76728	0.38624	1.01587	161	0	1	0.03	7.14	0
0	1	7.76454	0.12879	0.36111	157	0	1	0.06	10.02	0
0	0	7.61944	0.00000	0.57143	104	1	1	0.03	6.93	0
1	1	7.35113	0.13603	0.59829	152	1	1	0.03	8.80	0
0	1	6.99795	0.14290	0.66598	56	0	1	0.03	6.76	0
1	0	6.89938	0.00000	0.67747	150	1	1	0.03	7.75	0
1	0	6.83094	0.00000	0.60102	11	0	1	0.11	10.99	0
0	0	6.47775	0.00000	0.62208	72	0	1	0.03	8.23	0
0	0	6.02053	0.00000	1.22172	99	1	0	0.03	12.50	0
0	1	6.25873	0.15978	0.71778	79	1	1	0.03	11.74	0
0	0	6.17112	0.00000	0.88183	136	1	1	0.03	9.30	0
0	0	5.86995	0.00000	0.42159	117	1	1	0.03	8.90	0
0	0	5.73854	0.00000	0.67518	128	1	1	0.03	10.40	0
1	0	5.56331	0.00000	0.79498	129	0	1	0.03	7.30	0
0	0	5.52498	0.00000	0.40040	56	0	1	0.27	8.00	0
1	15	5.19918	2.88507	0.34169	176	0	1	0.03	8.60	0
1	14	5.31143	2.63582	0.54701	149	1	1	0.03	8.10	0
0	1	5.12799	0.19501	0.67935	57	0	1	0.04	7.10	1
1	3	5.04860	0.59422	0.70661	155	1	1	0.03	8.50	0
0	0	9.62081	0.00000	0.33635	50	0	1	0.20	10.65	0
0	8	9.62081	0.83153	0.81505	176	1	1	0.01	7.44	0
0	6	9.48665	0.63247	1.04405	84	1	0	0.01	8.96	0
1	27	9.48665	2.84610	1.53061	66	1	0	0.01	9.56	0
1	2	2.59274	0.77138	0.41176	157	0	1	0.07	9.11	0
0	0	9.39083	0.00000	0.51988	124	1	1	0.09	10.23	0
1	17	9.34155	1.81983	0.76823	71	0	1	0.12	7.01	0
0	0	9.31964	0.00000	0.46642	48	1	1	0.05	12.68	0
1	0	9.29500	0.00000	0.72488	59	0	0	0.01	9.35	1
1	2	8.01916	0.24940	0.45147	121	0	1	0.03	9.69	0
0	0	7.85763	0.00000	0.48544	30	0	1	0.03	8.98	0

etc.

doubly homogeneous Poisson model, the large sample test of H_0 in (8.20) yields

$$Z = \frac{0.6631 - 0.2189}{\sqrt{\frac{2266}{(2598.5 \times 2480.2)}}} = 23.689,$$

that is highly statistically significant. Likewise, the test based on the log relative risk in (8.25) yields $Z = 1.108/0.0492 = 22.52$.

8.1.4 Violations of the Homogeneous Poisson Assumptions

All of the above is based on the doubly homogeneous Poisson assumptions that the intensity of the process is constant over time, and that it is the same for all subjects in the population. These assumptions may not apply in practice. Violation of the first assumption is difficult, if not impossible, to assess with count data when the exact times of the events are not recorded. However, when the event times are known, a time-varying intensity is easily assessed or allowed for in a multiplicative intensity model that is a generalization of the Cox proportional hazards model for recurrent event times (see Section 9.7). The second assumption specifies that for given exposure time t , the mean number of events in the population is $E(d) = \lambda t$ and that the variance of the number of events also is $V(d) = \lambda t$, as is characteristic of the Poisson distribution. In any population, violation of either the homogeneous mean assumption or of the mean–variance relationship leads to over- or under-dispersion in the data, where $V(d) > (<) \lambda t$.

Cochran (1954a) suggested that the homogeneous Poisson process assumption be tested using a simple chi-square goodness-of-fit test. Under the assumption of a common intensity for all subjects, then $E(d_j) = V(d_j) = \lambda t_j$ and the test is provided by

$$X^2 = \sum_{j=1}^N \frac{(d_j - \hat{\lambda}t_j)^2}{\hat{\lambda}t_j}, \quad (8.28)$$

which is distributed as chi-square on $N - 1$ df under the hypothesis of homogeneity. An alternative approach is to assess the degree of over- or under-dispersion in a Poisson regression model, as shown in Section 8.4.

Example 8.2 Hypoglycemia in the DCCT (continued)

Applying the above to each DCCT treatment group separately yields chi-square values of $X^2 = 1737.12$ in the conventional treatment group on 351 df ($p \leq 0.001$) and 3394.22 in the intensive group on 362 df ($p \leq 0.001$). In each case there is a strong indication that the simple homogeneous Poisson model does not apply. These tests can also be obtained from a Poisson regression model as described in Example 8.4.

8.2 OVERDISPersed POISSON MODEL

One way to allow for this extra-variation in risks is to assume that the overdispersion arises because of mixtures of subjects with different characteristics, that, in turn, determine the intensity or risk of events in the population. This implies that the intensity for a subject is a function of covariate values and that all subjects with a given covariate vector \mathbf{x} share a common intensity $\lambda(\mathbf{x})$. These relationships can then be assessed using a Poisson regression model.

Another approach is to assume a random effects overdispersed Poisson model in which we assume that each subject has a unique intensity that is drawn from a distribution of intensities. We first consider this overdispersed model, followed by use of the Poisson regression model.

8.2.1 Two-Stage Random Effects Model

The doubly homogeneous Poisson model assumes that the mean and variance of the number of events for the j th subject with exposure t_j is $E(d_j|t_j) = V(d_j|t_j) = \lambda t_j$. Over- or under-dispersion arises when these relationships are violated. In this case it is more realistic to describe the process in terms of a two-stage random effects model, where the j th subject has a unique intensity λ_j that determines the mean and variance for that subject, and where the subject-specific intensities are then drawn from some distribution in the population. That is, we assume that

$$\begin{aligned} E(d|\lambda, t) &= V(d|\lambda, t) = \lambda t \\ \lambda &\sim G(\mu_\lambda, \sigma_\lambda^2) \end{aligned} \quad (8.29)$$

for some "mixing" distribution G , where $E(\lambda) = \mu_\lambda$ is now the overall mean rate in the population and $V(\lambda) = \sigma_\lambda^2$ is the *overdispersion variance component*. Throughout we also assume that λ is independent of t and we condition on the exposure times $\{t_j\}$ as fixed quantities.

Then unconditionally for the j th subject,

$$E(d_j) = E_\lambda E[d_j|\lambda_j, t_j] = \mu_\lambda t_j. \quad (8.30)$$

The rate of events for each subject then provides an estimate of the subject-specific rate

$$\hat{\lambda}_j = r_j = d_j/t_j, \quad (8.31)$$

where

$$\begin{aligned} E\left[\hat{\lambda}_j|\lambda_j\right] &= E\left[\frac{d_j}{t_j}|\lambda_j\right] = \frac{\lambda_j t_j}{t_j} = \lambda_j \\ V\left[\hat{\lambda}_j|\lambda_j\right] &= V\left[\frac{d_j}{t_j}|\lambda_j\right] = \frac{1}{t_j^2} V(d_j) = \frac{\lambda_j t_j}{t_j^2} = \frac{\lambda_j}{t_j}. \end{aligned} \quad (8.32)$$

Note that this model is directly analogous to the simple measurement error model described in Section 4.10.1. Here the subject-specific number of events is equivalent

to the subject-specific cholesterol value ($d_j \equiv y_j$), the subject-specific intensity is equivalent to the subject-specific true cholesterol value ($\lambda_j \equiv v_j$), the mean intensity is equivalent to the mean cholesterol ($\mu_\lambda \equiv \mu$), and the variance of the intensity between subjects is equivalent to the variance of the true cholesterol measurements ($\sigma_\lambda^2 \equiv \sigma_\mu^2$).

Also note that a dispersion variance component value of $\sigma_\lambda^2 = 0$ implies that there is no overdispersion, in which case G is degenerate and $\lambda_j = \mu_\lambda$ for all subjects. In this case the homogeneous Poisson model applies.

Therefore, the objective is to estimate the dispersion variance component σ_λ^2 , and if nonzero, to then estimate μ_λ and $V(\hat{\mu}_\lambda)$ allowing for overdispersion. Various authors have proposed methods that are based on a parametric specification of the mixing distribution $G(\mu_\lambda, \sigma_\lambda^2)$, such as the gamma mixing distribution or the log normal, among others. Here we take a different approach wherein an estimator of the variance component is obtained using distribution-free or robust methods that do not require specification of the specific form of the mixing distribution.

Consider the weighted mean rate in (8.9). Then

$$E(\bar{r}) = \frac{\sum_j E(d_j)}{\sum_j t_j} = \frac{\mu_\lambda \sum_j t_j}{\sum_j t_j} = \mu_\lambda, \quad (8.33)$$

and the crude rate \bar{r} provides an unbiased estimate of μ_λ under the random effects model. We now require an estimate of $V(\bar{r})$. Unconditionally, for the j th subject, the variance of d_j can be expressed as

$$\begin{aligned} V(d_j) &= V[E(d_j | \lambda_j t_j)] + E[V(d_j | \lambda_j t_j)] \\ &= V[\lambda_j t_j] + E[\lambda_j t_j] \\ &= t_j^2 \sigma_\lambda^2 + \mu_\lambda t_j, \end{aligned} \quad (8.34)$$

which reflects both the variance of the intensities between subjects and the Poisson variation within subjects. Thus,

$$V(\bar{r}) = \frac{\sum_j V(d_j)}{\left(\sum_j t_j\right)^2} = \frac{\sum_j t_j^2 \sigma_\lambda^2 + \mu_\lambda \sum_j t_j}{\left(\sum_j t_j\right)^2}. \quad (8.35)$$

The two-stage model described in Section 4.10 can then be employed to obtain a moment estimator for the dispersion parameter, σ_λ^2 , based on the expected value of a sum of squares that involves σ_λ^2 . Since $\hat{\mu}_\lambda = \bar{r}$ is a weighted mean of the $\{r_j\}$ with weights $\{t_j\}$, consider the weighted SSE. From (8.34),

$$E \left[\sum_j t_j (r_j - \mu_\lambda)^2 \right] = \sum_j t_j V(r_j) = \sum_j \frac{V(d_j)}{t_j} = \left(\sum_j t_j \sigma_\lambda^2 \right) + N \mu_\lambda. \quad (8.36)$$

Solving for σ_λ^2 we obtain the moment estimator

$$\hat{\sigma}_\lambda^2 = \max \left[0, \frac{\sum_j t_j (r_j - \hat{\mu}_\lambda)^2 - N \hat{\mu}_\lambda}{\sum_j t_j} \right]. \quad (8.37)$$

Substituting into the expression for $V(\bar{r})$ in (8.35), we obtain the estimated variance of the estimated mean intensity

$$\widehat{V}(\bar{r}) = \frac{\sum_j t_j^2 \widehat{\sigma}_\lambda^2 + \bar{r} \sum_j t_j}{\left(\sum_j t_j\right)^2}. \quad (8.38)$$

For confidence intervals it is customary that we use the log scale. Again let $\eta = \log(\mu_\lambda)$ and $\widehat{\eta} = \log(\bar{r})$, where

$$V(\widehat{\eta}) = \frac{V(\bar{r})}{\mu_\lambda^2} \cong \frac{\widehat{V}(\bar{r})}{\bar{r}^2}. \quad (8.39)$$

Under weak conditions, using the Liapunov central limit theorem (Section A.2.1) it can be shown that $\sum_j d_j$ is asymptotically normally distributed so that asymptotically

$$\bar{r} \xrightarrow{d} \mathcal{N}[\mu_\lambda, V(\bar{r})], \quad (8.40)$$

and

$$\widehat{\eta} = \log(\bar{r}) \xrightarrow{d} \mathcal{N}[\log(\mu_\lambda), V(\bar{r})/\mu_\lambda^2]. \quad (8.41)$$

Therefore, asymmetric confidence limits on the mean rate are obtained from the symmetric limits on the log mean rate using $\widehat{\eta} \pm Z_{1-\alpha/2} \sqrt{\widehat{V}(\widehat{\eta})}$.

A fixed-point iterative method can then be used to obtain jointly convergent estimates of the mean rate μ_λ and the overdispersion variance component σ_λ^2 . Let the initial values for the iteration be the values of the crude rate $\widehat{\mu}_\lambda^{(0)} = \bar{r}$ from (8.9) and the variance estimate $\widehat{\sigma}_\lambda^{2(0)}$ from (8.37). Then using the fixed-point method, we can alternatively compute updated estimates of the mean and the variance component until both converge. The updated estimate of the mean at the $(\ell + 1)$ th iteration ($\ell = 0, 1, 2, \dots$) can be obtained as an inverse variance weighted estimate

$$\widehat{\mu}_\lambda^{(\ell+1)} = \bar{r}^{(\ell+1)} = \frac{\sum_j \widehat{\tau}_j^{(\ell)} r_j}{\sum_j \widehat{\tau}_j^{(\ell)}}, \quad (8.42)$$

where $\widehat{\tau}_j^{(\ell)} = [\widehat{V}(r_j)^{(\ell)}]^{-1}$ and where, from (8.34),

$$\widehat{V}(r_j)^{(\ell)} = \frac{t_j^2 \widehat{\sigma}_\lambda^{2(\ell)} + \widehat{\mu}_\lambda^{(\ell)} t_j}{t_j^2} = \widehat{\sigma}_\lambda^{2(\ell)} + \frac{\widehat{\mu}_\lambda^{(\ell)}}{t_j}. \quad (8.43)$$

Then the updated estimate of the variance component $\widehat{\sigma}_\lambda^{2(\ell+1)}$ is obtained upon substitution of $\widehat{\mu}_\lambda^{(\ell+1)}$ into the moment estimating equation (8.37). The process continues until both the mean and variance estimates converge to constants.

8.2.2 Relative Risks

Again, consider the case of samples of n_1 and n_2 observations from two populations where within each we assume a two-stage model with mixing distributions $G_1(\mu_{\lambda_1}, \sigma_{\lambda_1}^2)$ and $G_2(\mu_{\lambda_2}, \sigma_{\lambda_2}^2)$, respectively. We then wish to estimate the relative risk, or its logarithm $\theta = \log(RR) = \eta_1 - \eta_2$, where $\eta_i = \log(\mu_{\lambda_i})$ and $\hat{\theta} = \hat{\eta}_1 - \hat{\eta}_2$. In this case, we first obtain an estimate of the dispersion variance components $\hat{\sigma}_{\lambda_1}^2$ and $\hat{\sigma}_{\lambda_2}^2$ separately within each group. These are then employed in (8.38) to obtain estimates of the variances $\hat{V}(\bar{r}_i)$ and in (8.39) to obtain estimates of the $\hat{V}(\hat{\eta}_i)$, $i = 1, 2$. The estimated variance of the estimated $\log(RR)$ then is $V(\hat{\theta}) = V(\hat{\eta}_1) + V(\hat{\eta}_2)$, which can be used as the basis for a large sample confidence interval calculation.

Now consider a test of the null hypothesis $H_0: \mu_{\lambda_1} = \mu_{\lambda_2}$. Ideally, we wish to construct a test using the variance under this null hypothesis. However, to do so we must also allow for the dispersion variance components to differ between groups, that is, assuming that $\sigma_{\lambda_1}^2 \neq \sigma_{\lambda_2}^2$. Thus, we first estimate μ_{λ} from the pooled sample with both groups combined. We then estimate $\sigma_{\lambda_i}^2$ separately for the i th group but using $\hat{\mu}_{\lambda}$ in the moment-estimating equation (8.37). Designate the result as $\hat{\sigma}_{\lambda_i(0)}^2$ to indicate that the dispersion variance is estimated under the null hypothesis. We would then substitute $\hat{\mu}_{\lambda}$ and $\hat{\sigma}_{\lambda_i(0)}^2$ in (8.38) to obtain an estimate of the $\hat{V}(\bar{r}_i|H_0)$ separately for each group ($i = 1, 2$). Then an asymptotically efficient test of H_0 is provided by

$$Z = \frac{(\bar{r}_1 - \bar{r}_2)}{\sqrt{\hat{V}(\bar{r}_1|H_0) + \hat{V}(\bar{r}_2|H_0)}}, \quad (8.44)$$

which is asymptotically normally distributed. Note that this test is also asymptotically efficient when there is homogeneity of dispersion parameters ($\sigma_{\lambda_1}^2 = \sigma_{\lambda_2}^2$) because the estimates within each group each provide a consistent estimate of the common parameter in this case.

Example 8.3 Hypoglycemia in the DCCT (continued)

The following is a summary of the overdispersed analysis of the rate of hypoglycemia within the two groups in the DCCT:

	Intensive $i = 1$	Conventional $i = 2$
$\sum_{j=1}^{n_i} t_{ij} (r_{ij} - \hat{\mu}_{\lambda_i})^2$	2250.64	380.309
$\hat{\sigma}_{\lambda_i}^2$	0.77351	0.12226
$\sum_{ij} t_{ij}^2$	18315.17	19488.77
$\hat{V}(\bar{r}_i)$	0.0024878	0.0004523
$\hat{V}(\hat{\eta}_i)$	0.0056582	0.0094363

These yield an estimated variance $\hat{V}(\hat{\theta}) = 0.0056582 + 0.0094363 = 0.015095$ and $S.E.(\hat{\theta}) = \sqrt{0.015095} = 0.12286$. The resulting asymmetric 95% confidence limits on RR are (2.381, 3.853), slightly wider than the homogeneous model confidence

limits presented in Example 8.1. Using this *S.E.* computed under the alternative yields a Wald *Z*-test of $Z = 1.10814/0.12286 = 9.01954$, still highly statistically significant.

In the combined sample, as shown in Example 8.1, the mean rate is $\hat{\mu}_\lambda = 0.44618$. This can then be used in (8.37) to compute the estimate of the dispersion variance component within each group under the null hypothesis of no difference in rates between groups, designated as $\hat{\sigma}_{\lambda_{i(0)}}^2$. When both quantities are employed in (8.38) this yields estimates of the variances of the rates within each group under the null hypothesis, $\hat{V}(\bar{r}_i|H_0)$. The resulting values for the dispersion variance components and the variances of the rates are

	<i>Intensive</i> $i = 1$	<i>Conventional</i> $i = 2$
$\hat{\sigma}_{\lambda_{i(0)}}^2$	0.85086	0.14165
$\hat{V}(\bar{r}_i H_0)$	0.0026276	0.0006016

Then the test of the difference between groups allowing for overdispersion is

$$Z = \frac{0.6631 - 0.2189}{\sqrt{0.0026276 + 0.0006016}} = 7.816,$$

which is again highly significant.

8.2.3 Stratified-Adjusted Analyses

As was the case for the analysis of odds ratios and relative risks of single binary events, in some cases it is desirable to conduct an analysis that is stratified or adjusted for other patient covariates. Such analyses are readily performed by adapting the methods described in Chapter 4 to the analysis of crude rates, either under a homogeneous Poisson model or under an overdispersed Poisson model with an additional parameter (the overdispersion variance component) estimated separately within each group and within each stratum. These methods include the *MVLE* of an assumed common relative risk over strata and its large sample variance, the Radhakrishna-like asymptotically fully efficient test, the Cochran test of homogeneity of relative risks over stratum, and the random effects stratified-adjusted estimate. These developments are left as problems.

8.3 POISSON REGRESSION MODEL

8.3.1 Homogeneous Poisson Regression Model

Another mechanism by which overdispersion may arise in the aggregate sample is when the population consists of a mixture of subpopulations characterized by different covariate values that, in turn, are associated with different intensities or risks. Thus, the population consists of a mixture of subjects with different covariate values that

implies a mixture of some subjects with inherently higher risks and some subjects with inherently lower risks. One way to account for such variation in risks is through a Poisson regression model in which the conditional expectation (the intensity) is modeled as a function of covariates (Frome, 1983). Since the Poisson distribution is a member of the exponential family, the Poisson regression model is a member of the family of generalized linear models (*GLMs*) described in Section A.10.

Assume that each patient ($i = 1, \dots, N$) has an underlying risk or intensity that is a function of a covariate vector $\mathbf{x}_i = (x_{i1} \ \dots \ x_{ip})^T$. Because the rate parameter $\lambda(\mathbf{x}_i)$ must be positive, it is natural to use a log-linear model with a log link which is the canonical link for the Poisson rate parameter. Thus,

$$\begin{aligned}\log[\lambda(\mathbf{x}_i)] &= (\alpha + x_{i1}\beta_1 + \dots + x_{ip}\beta_p) = \alpha + \mathbf{x}'_i \boldsymbol{\beta} = \eta_i \\ \lambda(\mathbf{x}_i) &= e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}}\end{aligned}\quad (8.45)$$

with parameters $\boldsymbol{\theta} = (\alpha \parallel \boldsymbol{\beta}^T)^T$, where $\boldsymbol{\beta} = (\beta_1 \ \dots \ \beta_p)^T$. For the j th covariate, β_j is the log relative risk per unit change in X_j . Then for the i th patient with d_i events over t_i years of exposure, the expected number of events is

$$E(d_i|\mathbf{x}_i, t_i) = \lambda(\mathbf{x}_i) t_i = \exp[\alpha + \mathbf{x}'_i \boldsymbol{\beta} + \log(t_i)]. \quad (8.46)$$

Here $\log(t_i)$ is an *offset*, so that each patient has an implied intercept $\tilde{\alpha}_i = \alpha + \log(t_i)$ given the fixed exposure time t_i .

Then assume that

$$d_i \sim \text{Poisson}[\lambda(\mathbf{x}_i) t_i] = P_o(d_i; \lambda(\mathbf{x}_i), t_i) \quad (8.47)$$

or that $E(d_i|\lambda(\mathbf{x}_i), t_i) = V(d_i|\lambda(\mathbf{x}_i), t_i) = \lambda(\mathbf{x}_i) t_i$. Alternatively, this is equivalent to specifying that $d_i = \lambda(\mathbf{x}_i) t_i + \varepsilon_i$, where the errors are distributed with mean zero and variance $\lambda(\mathbf{x}_i) t_i$ independent of t_i . Therefore, the likelihood is the product of Poisson probabilities of the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \frac{e^{-\lambda(\mathbf{x}_i) t_i} [\lambda(\mathbf{x}_i) t_i]^{d_i}}{d_i!}, \quad (8.48)$$

with log likelihood

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^N [-\lambda(\mathbf{x}_i) t_i + d_i \log[\lambda(\mathbf{x}_i) t_i] - \log(d_i!)] \\ &= \sum_{i=1}^N \left[-e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} t_i + d_i [\alpha + \mathbf{x}'_i \boldsymbol{\beta} + \log(t_i)] - \log(d_i!) \right].\end{aligned}$$

The score vector is $\mathbf{U}(\boldsymbol{\theta}) = \left[U(\boldsymbol{\theta})_\alpha \ U(\boldsymbol{\theta})_{\beta_1} \ \dots \ U(\boldsymbol{\theta})_{\beta_p} \right]^T$ with elements

$$U(\boldsymbol{\theta})_\alpha = \sum_{i=1}^N \left[d_i - e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} t_i \right] = \sum_{i=1}^N [d_i - \lambda(\mathbf{x}_i) t_i], \quad (8.49)$$

and

$$U(\boldsymbol{\theta})_{\beta_j} = \sum_{i=1}^N x_{ij} \left[d_i - e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} t_i \right] = \sum_{i=1}^N x_{ij} [d_i - \lambda(\mathbf{x}_i) t_i] \quad (8.50)$$

for $j = 1, \dots, p$. Therefore, $\mathbf{U}(\boldsymbol{\theta})_{\beta_j} = \sum_{i=1}^N \mathbf{x}_i [d_i - \lambda(\mathbf{x}_i) t_i]$.

From $\mathbf{U}(\boldsymbol{\theta})_{\alpha}$ it follows that $d_{\bullet} = \sum_i d_i = \sum_i \hat{\lambda}_i t_i$. Also, for the j th covariate, given fixed values of the other covariates and a fixed exposure time t , β_j equals the difference in the log intensity (log risk) per unit difference in the value of the covariate X_j , and e^{β_j} equals the log relative intensity (risk) per unit difference.

It is then readily shown that the observed and expected information matrices have elements

$$\begin{aligned} \mathbf{i}(\boldsymbol{\theta})_{\alpha} &= \mathbf{I}(\boldsymbol{\theta})_{\alpha} = \sum_i e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} t_i = \sum_i \lambda_i(\mathbf{x}_i) t_i, \\ \mathbf{i}(\boldsymbol{\theta})_{\alpha, \beta_j} &= \mathbf{I}(\boldsymbol{\theta})_{\alpha, \beta_j} = \sum_i x_{ij} t_i e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} = \sum_i x_{ij} \lambda_i(\mathbf{x}_i) t_i, \\ \mathbf{i}(\boldsymbol{\theta})_{\beta_j} &= \mathbf{I}(\boldsymbol{\theta})_{\beta_j} = \sum_i x_{ij}^2 t_i e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} = \sum_i x_{ij}^2 \lambda_i(\mathbf{x}_i) t_i, \\ \mathbf{i}(\boldsymbol{\theta})_{\beta_j, \beta_k} &= \mathbf{I}(\boldsymbol{\theta})_{\beta_j, \beta_k} = \sum_i x_{ij} x_{ik} t_i e^{\alpha + \mathbf{x}'_i \boldsymbol{\beta}} = \sum_i x_{ij} x_{ik} \lambda_i(\mathbf{x}_i) t_i \end{aligned} \quad (8.51)$$

for $1 \leq j < k \leq p$. Therefore, $\mathbf{i}(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\hat{\boldsymbol{\theta}})$ with elements obtained by substituting $\hat{\boldsymbol{\theta}} = (\hat{\alpha} \ \hat{\beta}_1 \ \dots \ \hat{\beta}_p)^T$ into the above expressions. As for any member of the exponential (GLM) family, the elements of the information are weighted sums of the model-based conditional variances, such as where $\mathbf{I}(\boldsymbol{\theta})_{\beta_j} = \sum_i x_{ij}^2 V(d_i | \mathbf{x}_i, t_i)$.

The estimated information then provides likelihood ratio tests, Wald tests, and efficient score tests. As a problem we show that the score test for the coefficient of a single binary covariate for treatment group equals the simple Z -test in (8.25) derived under the homogeneous Poisson model.

The model is readily fit to a data set using the SAS procedure for generalized linear models, PROC GENMOD, as illustrated in the following example.

Example 8.4 Hypoglycemia in the DCCT (continued)

We now conduct a Poisson regression analysis of the incidence of episodes of severe hypoglycemia in the DCCT, adjusting for the other covariates in the data set presented in Table 8.1. These include *insulin*: the baseline daily insulin dose in units per kilogram body weight; *duration*: the number of months duration of diabetes on entry into the study in the range 12 to 180; *female*: an indicator variable for female (1) versus male (0); *adult*: an indicator variable for adult (1, ≥ 18 years) versus adolescent on entry; *bcval5*: the level of c-peptide on entry in picomoles/mL; *hbael*: the percent glycosylated hemoglobin (HbA_{1c} %); and *hxcoma*: an indicator variable for a history of coma and/or seizure prior to entry (1 = yes, 0 = no).

Since we wish to quantify the increased risk with intensive therapy, the treatment variable has been recoded as *conv* = 0 if intensive, 1 if conventional group. Since the latter category is used as the reference in PROC GENMOD, then the coefficient for *conv* is the log relative risk for intensive versus conventional therapy.

The HbA_{1c} is a measure of the overall level of blood glucose control. The lower the level of HbA_{1c} , the greater will be the risk of hypoglycemia that occurs when the blood glucose falls below the levels required to maintain consciousness. The basal c-peptide is a measure of the residual endogenous insulin secreted by the pancreas; the higher the value, the less demand there is for external (exogenous) insulin. A small

Table 8.2 PROC GENMOD program for analysis of the DCCT data.

```

data one; set DCCT;
lnyears = log(T);
conv = 1 - int;
proc genmod;
  model nevents =
    / dist = poisson link = log offset = lnyears;
TITLE1 'Poisson regression models of risk of hypoglycemia';
title2 'null model';

proc genmod; class conv;
  model nevents = conv
    / dist = poisson link = log offset = lnyears;
title2 'unadjusted treatment group (conv) effect';

proc genmod; class conv;
  model nevents = conv insulin duration female
                adult bcval5 hbael hxcoma
    / dist = poisson link = log offset = lnyears covb;
title2 'covariate adjusted treatment group (conv) effect';

```

level of endogenous insulin (c-peptide) may be more protective against hypoglycemia than virtually no endogenous insulin. The c-peptide level diminishes to zero over time. The total insulin dose may reflect past efforts to maintain low blood glucose values or may reflect the need for exogenous insulin to compensate for deficiencies in endogenous insulin. Patients with multiple prior episodes of hypoglycemia were excluded from entry, but a small fraction reported a history of one or two prior such episodes. The objective is to explore the effects of these covariates on the risk of hypoglycemia and to obtain an assessment of the treatment group effect after adjusting for these possible risk factors.

Table 8.2 presents the program statements using PROC GENMOD to fit a Poisson regression model to these data. Here the variable $nevents = d_j$. The results of the analyses are presented in Tables 8.3 to 8.6, with some extraneous information deleted.

The first model presented in Table 8.3 is the null or intercept-only model. This null model yields an estimated intercept of $\hat{\alpha} = -0.807 = \log(0.4462) = \log(\hat{\lambda})$, where $\hat{\lambda}$ is the overall crude rate computed as in (8.23). Also, the $S.E.(\hat{\alpha}) = 1/\sqrt{(d_{1.} + d_{2.})} = 0.0210$. The log-likelihood value is $\log[L(\alpha)] = 479.1508$, with a corresponding $-2 \log[L(\alpha)] = -958.3$. This is not the complete log likelihood since the computation does not include the negative constant $-3371.1 = \sum_i \log(d_i!)$ (computed separately). This constant would cancel from any likelihood ratio test computation comparing this model to any other model. For this model, the *deviance*

Table 8.3 PROC GENMOD Poisson regression analysis of the risk of hypoglycemia in the DCCT: The null model.

null model
The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	714	4520.8419	6.3317
Scaled Deviance	714	4520.8419	6.3317
Pearson Chi-Square	714	6457.7296	9.0444
Scaled Pearson X2	714	6457.7296	9.0444
Log Likelihood	.	479.1508	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-0.8070	0.0210	1475.8849	0.0001
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

$= 4520.84$, which is obtained as the difference between the $-2 \log[L(\alpha)]$ and that of a model that fits the data perfectly (see Section A.10.3), where the log likelihood for the latter is $\log[L(\lambda_1, \dots, \lambda_N)] = 2739.57$, also ignoring the large negative constant (computed separately). Thus, the *deviance* $= -2(479.1508 - 2739.57) = 4520.84$.

Since PROC GENMOD fits the family of *GLMs* using quasi-likelihood, it can allow for a dispersion or scale parameter, as described in Section A.10.4. Since no overdispersion parameter is included in the models in this example, then in each model the scale parameter is fixed at the value 1.0. However, the Pearson chi-square is $X^2 = 6457.73$ on 714 *df*. This is the same as Cochran's variance test of the hypothesis of no-overdispersion described in Section 8.1.4. The associated *p*-value is $p \leq 0.0001$, which indicates substantial extra-variation. Nevertheless, models assuming homogeneous Poisson variation are presented here for illustration. Models allowing for overdispersion are presented subsequently in Section 8.4.

The second model presented in Table 8.4 then includes the effect of treatment group alone. Since the model includes a class variable (*conv*) for intensive (0) versus conventional (1) therapy, the latter being the reference category, then $\hat{\alpha} = \log(\hat{\lambda}_{conv}) = \log(0.2189) = -1.5190$; and $\hat{\beta} = \log(\widehat{RR}) = \log(\widehat{\lambda}_{int}/\widehat{\lambda}_{conv}) = \log(3.029) = 1.1081$. Also, $S.E.(\hat{\beta}) = \sqrt{1/d_{1\bullet} + 1/d_{2\bullet}} = 0.0492$. The program

Table 8.4 PROC GENMOD Poisson regression analysis of the risk of hypoglycemia in the DCCT: unadjusted treatment group effect.

Unadjusted treatment group effect
The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	713	3928.7828	5.5102
Scaled Deviance	713	3928.7828	5.5102
Pearson Chi-Square	713	5131.3432	7.1968
Scaled Pearson X2	713	5131.3432	7.1968
Log Likelihood	.	775.1804	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-1.5190	0.0429	1252.8961	0.0001
CONV 0	1	1.1081	0.0492	507.0072	0.0001
CONV 1	0	0.0000	0.0000	.	.
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

also computes the Wald test for each coefficient, which for the group effect yields a chi-square value of 507.0072 that corresponds to a Z -value of 22.517. All of these quantities equal those obtained from the homogeneous Poisson model analysis presented in Example 8.1. When contrasted to the *deviance* of the null model, this model also yields a likelihood ratio chi-square test of $4520.84 - 3928.78 = 592.06$, which corresponds to a Z -value of 24.33. These values are similar to the Z -value computed in Example 8.1 under homogeneous Poisson model assumptions.

The third model presented in Table 8.5 describes the effect of treatment group in addition to those of other covariates. The likelihood ratio test for the model is obtained as the difference in the deviance from the null model, so that $X^2_{LR} = 4520.84 - 3707.7027 = 813.14$ on $714 - 706 = 8$ *df*, which is highly significant. In this model, the adjusted $\widehat{\beta} = \log(\widehat{RR}) = 1.0845$ corresponds to an estimated $\widehat{RR} = 2.96$ adjusted for other covariates, virtually unchanged from the unadjusted analysis. Each of the other covariates, other than the levels of insulin and c-peptide (*bcval5*) were statistically significantly associated with the risk of hypoglycemia (based on Wald tests).

Table 8.5 PROC GENMOD Poisson regression analysis of the risk of hypoglycemia in the DCCT: treatment group adjusted for other covariate effects.

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	706	3707.7027	5.2517
Scaled Deviance	706	3707.7027	5.2517
Pearson Chi-Square	706	4792.7878	6.7887
Scaled Pearson X2	706	4792.7878	6.7887
Log Likelihood	.	885.7204	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-0.9568	0.2174	19.3765	0.0001
CONV 0	1	1.0845	0.0493	483.9072	0.0001
CONV 1	0	0.0000	0.0000	.	.
INSULIN	1	0.0051	0.0995	0.0026	0.9593
DURATION	1	0.0015	0.0006	6.7882	0.0092
FEMALE	1	0.1794	0.0424	17.9276	0.0001
ADULT	1	-0.5980	0.0656	83.1309	0.0001
BCVAL5	1	-0.5283	0.3630	2.1175	0.1456
HBAEL	1	-0.0335	0.0151	4.8868	0.0271
HXCOMA	1	0.6010	0.0685	77.0919	0.0001
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

As described in Section 7.3.4, PROC GENMOD will also compute likelihood ratio tests for each of the covariates in the model using the *type3* option. Such tests could have been computed here as well.

The covariance of the estimates is also presented in Table 8.6. These allow estimation of the variance of the estimated rates for individual subjects with specific covariate values. For example, for the first patient in the listing in Table 8.1, the linear predictor is $\hat{\eta}_i = \hat{\alpha} + \mathbf{x}'_i \hat{\beta} = -0.63870$, which yields an estimated rate per year of $\hat{\lambda}(\mathbf{x}_i) = \exp(-0.6387) = 0.52798$. Given the $t_i = 9.50582$ years of follow-up, this yields a predicted number of events of $\hat{\lambda}(\mathbf{x}_i)t_i = 5.0189$. Using the above covariance matrix, the *S.E.* of the linear predictor is 0.0537, which yields 95%

Table 8.6 PROC GENMOD Poisson regression analysis of the risk of hypoglycemia in the DCCT: estimated covariance matrix in the adjusted model.

Estimated Covariance Matrix						
Parameter	INTERCEPT	CONV	INSULIN	DURATION	FEMALE	
INTERCEPT	0.04725	-0.001773	-0.01189	-0.000051	-0.000241	
CONV	-0.001773	0.002431	0.0000531	-5.444E-9	0.000021	
INSULIN	-0.01189	0.0000531	0.009890	-8.33E-7	-0.000107	
DURATION	-0.000051	-5.444E-9	-8.33E-7	3.1565E-7	-2.211E-6	
FEMALE	-0.000241	0.0000214	-0.000107	-2.211E-6	0.001795	
ADULT	-0.006857	0.0000547	0.003462	-6.594E-6	0.000089	
BCVAL5	-0.02945	0.0004454	0.006824	0.0000956	-0.000535	
HBAEL	-0.002572	-0.00002	0.0001494	2.0192E-6	-0.000048	
HXCOMA	-0.001927	-0.000127	-0.000069	2.6715E-6	0.000013	
	ADULT	BCVAL5	HBAEL	HXCOMA		
	-0.006857	-0.02945	-0.002572	-0.001927		
	0.0000547	0.0004454	-0.00002	-0.000127		
	0.003462	0.006824	0.0001494	-0.000069		
	-6.594E-6	0.0000956	2.0192E-6	2.6715E-6		
	0.000089	-0.000535	-0.000048	0.000013		
	0.004301	-0.001021	0.0001582	0.0000196		
	-0.001021	0.13179	0.0008153	0.002335		
	0.0001582	0.0008153	0.0002291	0.0001239		
	0.0000196	0.002335	0.0001239	0.004686		

confidence limits of $(-0.7441, -0.5333)$. This yields 95% limits on the predicted number of events of $(4.5169, 5.5766)$. The linear predictor $\hat{\eta}_i$ and its *S.E.* may also be obtained from PROC GENMOD using the statements

```
%global _disk_ ; %let _disk_=on ;
%global _print_ ; %let _print_=off ;
PROC GENMOD;
MAKE 'OBSTATS' OUT=PRED ;
model nevents = conv insulin duration female
               adult bcval5 hbael hxcoma
/ dist = poisson link = log offset = lnyears covb OBSTATS;
PROC PRINT DATA=PRED;
```

8.3.2 Explained Variation

Using the developments in Section A.8.1, the explained fraction of squared error loss from a Poisson regression model is readily shown to be

$$\hat{R}_{\varepsilon^2}^2 = \frac{\hat{V}[E(d|\mathbf{x}, t)]}{\hat{V}(d|t)} = \frac{\sum_i [\hat{\lambda}(\mathbf{x}_i) t_i - \hat{\lambda} t_i]^2}{\sum_i [d_i - \hat{\lambda} t_i]^2} = R_{\varepsilon^2, resid}^2, \quad (8.52)$$

which equals the fraction of explained residual variation. In this expression, $\hat{\lambda} = \bar{r}$ is the crude rate estimate in the combined sample that also equals $e^{\hat{\alpha}}$ from the null model. The derivations are left to a problem.

Alternatively, the fraction of $-\log$ likelihood explained R_{ℓ}^2 can be used as described in Section A.8.3, or Maddala's measure R_{LR}^2 based on the likelihood ratio statistic as in Section A.8.4.

Example 8.5 Hypoglycemia in the DCCT (continued)

Using the output of the data set PRED, with additional calculations, it is then possible to compute the sums of squares in the numerator and denominator of (8.52). The resulting value of the $R_{\varepsilon^2, resid}^2 = (3139.217/21856.4) = 0.14363$.

Alternatively, the Madalla's likelihood ratio measure yields $R = R_{LR}^2 = 1 - \exp(-813.14/715) = 0.6793$, which is clearly substantially different from the fraction of squared error loss explained by the model.

The other possible measure of explained variation is the fraction of explained negative log likelihood. As described in Example 8.4, PROC GENMOD only computes the log likelihood and the deviance up to an additive constant. Thus, it is necessary to compute the complete log likelihood using a separate program based on the output data set (PRED) described above. The complete log likelihoods are $\log[L(\alpha)] = -2891.94$ and $\log[L(\alpha, \beta)] = -2485.37$, for which $R_{\ell}^2 = (2891.94 - 2485.37)/2891.94 = 0.14059$, which is rather close to the value of the $R_{\varepsilon^2, resid}^2$.

For illustration, the value of the ratio based on the kernel of the log likelihoods using the values presented in Tables 8.3 and 8.5, which ignore the constant, is $(885.7204 - 479.1508)/479.1508 = 0.84852$. This is clearly incorrect.

Alternatively, one could use the fraction of the explained deviance as a measure of explained variation, in which case the constant is irrelevant because it cancels in the computation of the deviance. This yields $R_D^2 = (4520.8419 - 3707.7027)/4520.8419 = 0.17986$, which is more in line with the value of R_{ℓ}^2 .

8.3.3 Applications of Poisson Regression

Example 8.5 illustrates the application of Poisson regression to the analysis of subject or unit-specific count data. Count data also arise in other contexts, such as the analysis of rates of events within subpopulations or vital statistics. An example is given in Problem 8.6. Another example is presented by Gail (1978). Poisson regression has

also been applied to the analysis of grouped survival data by Holford (1980) and Laird and Oliver (1981). In each instance the data structure consists of a count of the number of events associated with a measure of exposure, such as population size, within subsets of the population, where each subset is characterized by covariate values. The risk (rate) within each subpopulation is then expressed as a function of the covariate values.

8.4 OVERDISPersed AND ROBUST POISSON REGRESSION

The ordinary Poisson regression model, and the above analyses of the DCCT, all assume that there is no overdispersion after allowing for covariates. That is, the model assumes that the observed number of events for all subjects with a given covariate vector \mathbf{x} and fixed time t , or the conditional distribution $d|\mathbf{x}$, is distributed as Poisson with parameter $\lambda(\mathbf{x})$. However, even after adjusting for covariates, the conditional distribution of $d|\mathbf{x}$ may be overdispersed. There is a strong suggestion that this also applies to the analysis of hypoglycemia in the DCCT. The null model Pearson chi-square value shown in Table 8.3, which is equivalent to the Cochran variance test, is $X^2 = 6457.73$ on 714 df with $p \leq 0.0001$, indicating a significant degree of extra-variation. Another indication is that the $S.E.(\hat{\beta})$ for the effect of treatment group in the above covariate-adjusted regression model (0.0493) is still too small compared to the variance of the $\log(\bar{R})$ estimated from the two-stage overdispersed Poisson model (0.12286). Here we consider generalizations that allow for various kinds of model misspecification such as overdispersion.

8.4.1 Quasi-likelihood Overdispersed Poisson Regression

Many methods have been described for fitting overdispersed Poisson regression models. Breslow (1984, 1990) employed the method of moments to estimate the overdispersion variance component that is assumed to be homoscedastic for all values of the covariates. He then used quasi-likelihood to estimate the regression coefficients. Moore (1986) also discusses this approach. Others adopt a specific form for the mixing distribution, such as the Gaussian by Dean and Lawless (1989). An alternative approach is to adopt a quasi-likelihood that incorporates a scale or variance inflation dispersion parameter as described in Section A.10.3 that is also implemented in PROC GENMOD.

The overdispersed quasi-likelihood Poisson errors regression model adopts the simplifying assumption that

$$V(d_i|\mathbf{x}_i) = \nu E(d_i|\mathbf{x}_i, t_i) = \nu \lambda(\mathbf{x}_i) t_i \quad (8.53)$$

for scale parameter $\nu \neq 1$, or that the true conditional variance is ν times the expected variance (the mean) under a homogeneous Poisson model. As described in Section A.10.3, both the deviance and the Pearson chi-square tests of goodness of fit are asymptotically distributed as chi-square on $N - p - 1$ df when the model

Table 8.7 PROC GENMOD overdispersed Poisson regression analysis of the DCCT data using *Pearson/df*.

The GENMOD Procedure					
Criteria For Assessing Goodness Of Fit					
Criterion	DF	Value	Value/DF		
Deviance	706	3707.7027	5.2517		
Scaled Deviance	706	546.1619	0.7736		
Pearson Chi-Square	706	4792.7878	6.7887		
Scaled Pearson X2	706	706.0000	1.0000		
Log Likelihood	.	130.4707	.		

Analysis Of Parameter Estimates					
Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-0.9568	0.5664	2.8542	0.0911
CONV 0	1	1.0845	0.1285	71.2818	0.0001
CONV 1	0	0.0000	0.0000	.	.
INSULIN	1	0.0051	0.2591	0.0004	0.9844
DURATION	1	0.0015	0.0015	0.9999	0.3173
FEMALE	1	0.1794	0.1104	2.6408	0.1042
ADULT	1	-0.5980	0.1709	12.2456	0.0005
BCVAL5	1	-0.5283	0.9459	0.3119	0.5765
HBAEL	1	-0.0335	0.0394	0.7198	0.3962
HXCOMA	1	0.6010	0.1783	11.3560	0.0008
SCALE	0	2.6055	0.0000	.	.

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Squared/DOF.

assumptions are correct. Thus, a simple moment estimator of ν is provided by either $\hat{\nu} = \text{deviance}/df$ or by $\hat{\nu} = \text{Pearson } \chi^2/df$. Since the modes of convergence for the deviance and the Pearson statistics differ for some models (see McCullagh and Nelder, 1989) it is generally recommended that the Pearson statistic be used. When the model assumptions are correct, the 95% tolerance limits of the distribution of χ^2/df are $1 \pm 2.77/\sqrt{df} = 1 \pm 0.104$ for the covariate-adjusted Poisson regression model in Table 8.5 on 706 df. The Pearson $\chi^2/df = 6.7887$ thus indicates some degree of overdispersion. This provides the basis for specifying a starting value for an iterative procedure by which PROC GENMOD obtains an estimate of ν .

PROC GENMOD uses a different parameterization where the scale parameter is $scale = \sqrt{\nu}$. In the *model* statement we then use the option "scale=scale₀ pscale", where $scale_0 = \sqrt{\nu_0}$ is the starting value for the solution of $\hat{\nu}$. Therefore, in this example, $scale_0 = \sqrt{6.7887} = 2.606$ is the starting value. The solution $\hat{\nu}$ is that value such that the *scaled Pearson chi-square* for the fitted model equals

$$\frac{\text{Pearson } X^2 / df}{\widehat{\nu}} = 1.0. \quad (8.54)$$

All of the elements of $V(\widehat{\beta})$ are then multiplied by $\widehat{\nu}$, and all the $S.E.(\widehat{\beta})$ by $\sqrt{\widehat{\nu}}$. The $\widehat{\beta}$ themselves are unchanged.

Dscale is the corresponding option to estimate the scale parameter using *deviance/df*.

Example 8.6 Hypoglycemia in the DCCT (continued)

Table 8.7 presents the overdispersed covariate-adjusted model using the options in the *model* statement as follows:

```
/ dist = poisson link = log offset = logyears
  SCALE=2.606 PSCALE;
```

The scale parameter is estimated to be $\sqrt{\widehat{\nu}} = 2.6055$ as above and the standard errors of all the coefficients have been inflated by this amount. Thus, the $S.E.(\text{conv}) = (2.6055 \times 0.0493) = 0.1285$, which is comparable to that estimated from the random effects model (0.1229) without adjustment for covariates.

8.4.2 Robust Inference Using the Information Sandwich

Essentially, overdispersion arises when the variance of the random errors is misspecified. Another approach, therefore, is to base inferences on the robust information sandwich estimate of the variance–covariance matrix of the estimates that is based on the empirical estimate of the observed information. From the developments in Section A.9, it is readily shown that the empirical estimate of the covariance matrix of the Poisson model score vector is the matrix

$$\widehat{\mathbf{J}}(\widehat{\boldsymbol{\theta}}) = \sum_i \mathbf{U}(\widehat{\boldsymbol{\theta}}) \mathbf{U}(\widehat{\boldsymbol{\theta}})^T = \mathbf{X}' \widehat{\boldsymbol{\Sigma}}_{\varepsilon} \mathbf{X}, \quad (8.55)$$

where \mathbf{X} is the $n \times (p + 1)$ design matrix and $\widehat{\boldsymbol{\Sigma}}_{\varepsilon} = \text{diag}[(d_i - \widehat{\lambda}(\mathbf{x}_i)t_i)^2]$. When the scores, and thus the covariance matrix, are estimated under a specified null hypothesis, this matrix can be used as the basis for a robust efficient score test as described in the Appendix.

The robust information sandwich estimate of the covariance matrix of the errors is obtained as

$$\widehat{\boldsymbol{\Sigma}}_R(\widehat{\boldsymbol{\theta}}) = \mathbf{I}(\widehat{\boldsymbol{\theta}})' \widehat{\mathbf{J}}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{I}(\widehat{\boldsymbol{\theta}}), \quad (8.56)$$

where from (8.51) $\mathbf{I}(\widehat{\boldsymbol{\theta}}) = \mathbf{X}' \widehat{\boldsymbol{\Gamma}} \mathbf{X}$ and $\widehat{\boldsymbol{\Gamma}} = \text{diag}[\widehat{\lambda}(\mathbf{x}_i)t_i]$. This robust covariance matrix could be used as the basis for the computation of Wald tests and confidence limits.

Such computations can be performed using a supplemental SAS PROC IML program as described in Section 7.3.5 for a logistic regression model. Alternatively, computation of the robust information sandwich and associated confidence limits and Wald tests, but not the robust score test, can be obtained from PROC GENMOD.

Table 8.8 PROC GENMOD robust Poisson regression analysis of the DCCT data: Robust information sandwich covariance estimate.

The GENMOD Procedure						
Covariance Matrix (Empirical)						
Covariances are Above the Diagonal and Correlations are Below						
PRM1	PRM2	PRM4	PRM5	PRM6	PRM7	
(intercept	conv	insulin	duration	female	adult	
0.32773	-0.01447	-0.09258	-0.000257	0.005107	-0.05755	
-0.20799	0.01478	-0.000606	5.1525E-6	-0.001735	0.002650	
-0.52964	-0.01607	0.09303	-0.000043	-0.007640	0.03841	
-0.32059	0.03030	-0.10143	1.9571E-6	-7.34E-6	-0.00007	
0.07366	-0.11785	-0.20681	-0.04332	0.01467	-0.004818	
-0.47736	0.10353	0.59801	-0.23659	-0.18889	0.04435	
-0.50214	0.04150	0.17092	0.47185	0.10377	-0.03604	
-0.73454	0.07888	-0.02292	0.19724	-0.02465	2.4554E-6	
-0.20426	-0.01986	0.004646	0.23973	0.06898	0.009142	
PRM8	PRM9	PRM10				
bcval5	hbael	hxcoma)				
-0.28247	-0.01729	-0.02070				
0.004957	0.0003942	-0.000427				
0.05123	-0.000287	0.0002503				
0.0006486	0.0000113	0.0000594				
0.01235	-0.000123	0.001479				
-0.007458	2.1256E-8	0.0003409				
0.96557	0.01293	0.04096				
0.32006	0.001690	0.0007505				
0.23545	0.10312	0.03135				

One feature of PROC GENMOD is the option to conduct an analysis of correlated observations using generalized estimating equations (*GEE*), as described in Section A.10.6. *GEE* models employ the information sandwich to estimate the covariance matrix of the estimated coefficients when the observations may be correlated, such as where repeated measures are obtained on the same subject. This is designated by the *repeated* statement. When there is only a single observation per subject, as herein, the *repeated* statement may still be used in which case the robust information sandwich above is computed.

Table 8.9 PROC GENMOD robust Poisson regression analysis of the DCCT data: Robust analysis of parameter estimates.

Parameter	Analysis Of GEE Parameter Estimates						
	Empirical Standard Error Estimates						
	Estimate	Std Err	Lower	Upper	Z	Pr> z	
INTERCEPT	-0.9568	0.5725	-2.0789	0.1652	-1.671	0.0946	
CONV 0	1.0845	0.1216	0.8463	1.3228	8.9216	0.0000	
CONV 1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
INSULIN	0.0051	0.3050	-0.5927	0.6029	0.0166	0.9867	
DURATION	0.0015	0.0014	-0.0013	0.0042	1.0463	0.2954	
FEMALE	0.1794	0.1211	-0.0580	0.4168	1.4811	0.1386	
ADULT	-0.5980	0.2106	-1.0107	-0.1852	-2.839	0.0045	
BCVAL5	-0.5283	0.9826	-2.4542	1.3976	-0.5376	0.5908	
HBAEL	-0.0335	0.0411	-0.1140	0.0471	-0.8139	0.4157	
HXCOMA	0.6010	0.1770	0.2540	0.9480	3.3947	0.0007	
Scale	2.5891

NOTE: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's chi-square.

Example 8.7 Hypoglycemia in the DCCT (continued)

The following SAS statements would be used for the robust analysis of the DCCT hypoglycemia data employed in Table 8.2:

```
proc genmod; class conv patid;
model nevents = conv insulin duration female
adult bcval5 hbael hxcoma
/ dist = poisson link = log offset = logyears;
repeated subject=patid / type=unstr covb;
title2 'covariate adjusted treatment group effect';
```

Because there is only one observation per subject, indicated by the class effect *patid* in this case, then basically the *type=unstr* option specifies that the covariance matrix of the scores be estimated empirically. The robust (empirical) estimate of the covariance matrix of the coefficient estimates is presented in Table 8.8, and the robust analysis of the parameter estimates is presented in Table 8.9, extraneous information deleted. The coefficient estimates are unchanged from those shown previously because the score vectors are identical to those employed in the previous models, there being only one observation per subject. The robust information sandwich variance/covariance estimates are similar to those obtained from the quasi-likelihood analysis with a single scale or dispersion parameter.

8.4.3 Zeros-inflated Poisson Regression Model

Yet another approach to modeling overdispersed count data is to adopt a model that allows for extra zeros, or a count distribution with more zero frequencies than would be expected from a Poisson distribution. This can be viewed as a mixture of two subpopulations. In one, the subject is not at risk of an event owing to unique patient characteristics, in which case $d = 0$ regardless of the period of exposure. Let ω denote the fraction (probability) of subjects in this subpopulation. Then in the other subpopulation, the counts are distributed as Poisson with a probability that $d = 0$ equal to that specified by the Poisson distribution. Thus, the model assumes that

$$\begin{aligned}\Pr(d = 0) &= \omega + (1 - \omega)P_o(0; t, \lambda) \\ \Pr(d > 0) &= (1 - \omega)P_o(d; t, \lambda)\end{aligned}$$

so that the probability of a count $d > 1$ is provided by a rescaled Poisson distribution and that of $d = 0$ by a probability that is greater than that provided by a Poisson. This approach to modeling count data has a long history (see, e.g., Martin and Katti, 1965). Lambert (1992) was among the first to describe a model in which the probability of being immune to an event (ω) and the Poisson rate parameter (λ) were expressed as functions of covariates and the corresponding coefficients estimated using maximum likelihood estimation.

A zeros-added or zeros-inflated Poisson (ZIP) regression model then assumes that the probability ω_i for the i th subject is determined by a linear function of a covariate vector, say \mathbf{z}_i , with coefficient vector γ and link function $h(\omega_i) = \mathbf{z}'_i \gamma$, including the intercept. The link function $h(\cdot)$ can be any function appropriate for binary response data such as the logit. The inverse function yields $\omega(\mathbf{z}_i)$. The Poisson rate parameter λ_i is then expressed as before as a function of the covariate vector \mathbf{x}_i with coefficient vector β and link function $g(\lambda_i) = \mathbf{x}'_i \beta$, including the intercept, where the log link is the default. The vectors \mathbf{Z} and \mathbf{X} may be distinct covariate vectors, although they need not be. The likelihood is of the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \left[[1 - \omega(\mathbf{z}_i)] \frac{e^{-\lambda(\mathbf{x}_i)t_i} [\lambda(\mathbf{x}_i)t_i]^{d_i}}{d_i!} + \omega(\mathbf{z}_i)I[d_i = 0] \right], \quad (8.57)$$

that leads to direct estimation of the parameters $\boldsymbol{\theta} = (\beta, \gamma)$.

The model can be fit using PROC GENMOD with a *model* statement for the Poisson component, and a *zeromodel* statement for the added-zeros component. An example will clarify. However, since the ZIP model is fit using maximum likelihood rather than using a quasi-likelihood, it cannot be used in conjunction with either the *pscale* or *dscale* options that specify an overdispersion scale parameter as in Table 8.7; or with the *repeated* statement that would also provide a robust information sandwich covariance estimate as in Tables 8.8 and 8.9.

Example 8.8 Hypoglycemia in the DCCT (continued)

Table 8.2 presents the SAS statements to fit a model to the DCCT hypoglycemia data with no covariates, with the results presented in Table 8.3. A no-covariate ZIP model would be fit using the statements

Table 8.10 PROC GENMOD zeros-inflated Poisson (ZIP) regression analysis of the DCCT data, null (intercept only) model.

The GENMOD Procedure						
Criteria For Assessing Goodness Of Fit						
Criterion	DF	Value	Value/DF			
Deviance		4272.1581				
Scaled Deviance		4272.1581				
Pearson Chi-Square	713	1924.3076		2.6989		
Scaled Pearson X2	713	1924.3076		2.6989		
Log Likelihood		1235.0097				

Analysis Of Maximum Likelihood Parameter Estimates						
		Standard	Wald 95%	Wald	Pr >	
Parameter	DF	Estimate	Error	Confidence Limits	ChiSq	ChiSq
Intercept	1	-0.2655	0.0213	-0.3073 -0.2237	155.03	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000	

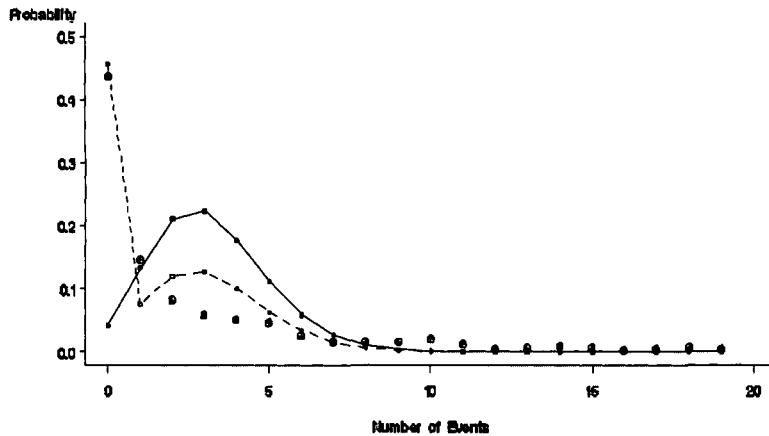
Maximum Likelihood Zero Inflation Parameter Estimates						
		Standard	Wald 95%	Wald	Pr >	
Parameter	DF	Estimate	Error	Confidence Limits	ChiSq	ChiSq
Intercept	1	-0.2723	0.0763	-0.4218 -0.1228	12.75	0.0004

```
proc genmod data=one;
  model nevents = / dist = zip link = log offset = lnyears covb;
  zeromodel / link=logit;
```

This model produces the computations in Table 8.10. There is been a substantial improvement in the deviance and Pearson chi-square compared to the no-covariate model in Table 8.3, the value/df of the latter having been reduced from 6.33 to 2.7. Whereas $\log(\hat{\lambda}) = -0.807$ in Table 8.3, allowing for the extra zeros yields an estimate of -0.2655 with corresponding estimates of $\hat{\lambda} = 0.446$ versus 0.767 . Figure 8.1 shows the resulting probabilities for counts ranging from 0 to 20. As assumed by the ZIP model, there is a lump of probability at a count of zero, and those for values 1 to 8 are then lower than those estimated from the homogeneous Poisson model.

The estimate of the logit of the probability ω of being resistant or immune to hypoglycemia is -0.2723 , so that $\hat{\omega} = (1 + e^{0.2723})^{-1} = 0.432$, demonstrating the high fraction of zeros beyond that expected from the Poisson distribution. In this cohort, 313 (43.8%) of the 715 subjects have a count of zero. With a rate parameter of $\hat{\lambda} = 0.767$, the probability of a zero over the mean follow-up of 7.1

Fig. 8.1 Poisson model (solid line), and the zeros-inflated Poisson (ZIP, -), fit to the overall DCCT hypoglycemia data. The empirical distribution is designated by \oplus .



years is $P_o(0, 0.767, 7.1) = e^{-7.1(0.767)} = 0.0043$, so that this model provides a better description of the overall distribution of events in this cohort.

The following SAS statements would fit a zeros-inflated Poisson (ZIP) model with covariates:

```
proc genmod data=one; class conv;
  model nevents = conv insulin duration female
    adult bcval5 hbael hxcoma
  / dist = zip link = log offset = lnyears;
  zeromodel conv insulin duration female
    adult bcval5 hbael hxcoma / link=logit;
```

The ZIP model results are presented in Table 8.11. Compared to the no-covariate ZIP model in Table 8.10, the addition of the covariates has resulted in a substantial decrease in the deviance (4272 to 3879) and in the Pearson chi-square; however, the value/df is changed slightly from 2.7 to 2.4.

The goodness-of-fit criteria for the simple Poisson model in Table 8.5 cannot be compared to the ZIP model because the no-covariate models are different. Also, parameter estimates in the ZIP model will differ from those of the simple Poisson model, owing to the reparameterization of the model, as seen above with the change in the intercept in the no-covariate model.

In this ZIP model with $\mathbf{X} = \mathbf{Z}$, each covariate has two effects. The coefficient in the Poisson component reflects a covariate effect on the log of the Poisson rate parameter, whereas that in the "zero inflation" section represents the log odds ratio of being immune to an event per unit change in the covariate (or between groups). Thus, it is possible that covariates will affect the Poisson rate and the extra-zero probability differentially. In Table 8.11, the intensive treatment group has a higher Poisson risk but a lower probability of extra zeros, and a history of coma (*hxcoma*) had a strong positive effect on the Poisson rate, but a strong negative effect on the probability of extra zeros. These are as expected since a factor that increases the risk of higher counts might be expected to lower the risk of no events.

However, in Figure 8.1, the empirical proportions for each count of events are also shown (as a \oplus). Clearly, the Poisson model does not come close to matching this empirical distribution. The zeros-added model is an improvement but still does not match the decreasing trend in the counts from 0 to 8 or so. This suggests that an entirely different type of model would be appropriate for these data. One such option is the negative binomial that is discussed in a later section.

8.5 CONDITIONAL POISSON REGRESSION FOR MATCHED SETS

As in Chapter 7, now consider the assessment of covariate effects on the intensity or rate parameter for individual subjects sampled in matched sets. For example, in a prospective study of the number of episodes of some condition (hypoglycemia in diabetes, epileptic seizures, etc.) subjects may be matched with respect to one or more covariates. As in Chapter 7, the data consist of N matched sets of size n_i for the i th set. The j th member of the i th set has covariate vector $\mathbf{x}_{ij} = (x_{ij1} \cdots x_{ijp})^T$ and the count variable d_{ij} designates the number of events experienced during exposure time t_{ij} for $i = 1, \dots, N$; $j = 1, \dots, n_i$. The Poisson model then specifies that the rate λ_{ij} for the ij th member is a function of the covariates \mathbf{x}_{ij} through the link function $g(\lambda_{ij}) = \alpha_i + \mathbf{x}'_{ij}\beta$. The model allows each matched set to have a unique risk of the outcome represented by a set-specific intercept α_i , and then assumes that the covariates have a common effect on the odds of the outcome over all matched sets represented by the coefficient vector $\beta = (\beta_1 \cdots \beta_p)^T$.

The unconditional Poisson regression model with parameters $\alpha = (\alpha_1 \cdots \alpha_N)^T$ and the vector of coefficients β then is of the form

$$L(\alpha, \beta) = \prod_{i=1}^N \prod_{j=1}^{n_i} \frac{e^{-\lambda(\mathbf{x}_{ij})t_{ij}} [\lambda(\mathbf{x}_{ij})t_{ij}]^{d_{ij}}}{d_{ij}!}, \quad (8.58)$$

where $\lambda(\mathbf{x}_{ij}) = \exp(\alpha_i + \mathbf{x}'_{ij}\beta)$. To fit this model, we must solve for the N nuisance parameters $\{\alpha_i\}$ in order to obtain an estimate of the coefficient vector β . Again, however, this can be avoided by conditioning on the appropriate sufficient statistic within each matched set, which in this case is the total number of events within the i th set, $d_{i\bullet} = \sum_j d_{ij}$.

Table 8.11 PROC GENMOD zeros-inflated Poisson (ZIP) regression analysis of the DCCT data.

The GENMOD Procedure							
Criteria For Assessing Goodness Of Fit							
Criterion	DF	Value		Value/DF			
Deviance		3879.9758					
Scaled Deviance		3879.9758					
Pearson Chi-Square	697	1686.8039		2.4201			
Scaled Pearson X2	697	1686.8039		2.4201			
Log Likelihood		1431.1008					
Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Confidence Limits	Wald ChiSq	Wald ChiSq	Pr >
Intercept	1	-0.9144	0.2252	-1.3557 -0.4731	16.49	16.49	<.0001
CONV 0	1	0.6767	0.0522	0.5744 0.7789	168.15	168.15	<.0001
CONV 1	0	0.0000	0.0000	0.0000 0.0000	.	.	.
insulin	1	-0.0603	0.1002	-0.2568 0.1362	0.36	0.36	0.5476
duration	1	0.0019	0.0006	0.0007 0.0031	10.23	10.23	<.0001
female	1	0.2134	0.0439	0.1274 0.2995	23.62	23.62	<.0001
adult	1	-0.4308	0.0660	-0.5601 -0.3014	42.60	42.60	<.0001
bcval5	1	0.4823	0.3874	-0.2770 1.2416	1.55	1.55	0.2131
hbael	1	0.0227	0.0159	-0.0084 0.0537	2.05	2.05	0.1527
hxcoma	1	0.3271	0.0694	0.1912 0.4631	22.24	22.24	<.0001
Scale	0	1.0000	0.0000	1.0000 1.0000			
Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Confidence Limits	Wald ChiSq	Wald ChiSq	Pr >
Intercept	1	-2.8171	0.9042	-4.5893 -1.0450	9.71	9.71	0.0018
CONV 0	1	-1.0725	0.1678	-1.4014 -0.7436	40.85	40.85	<.0001
CONV 1	0	0.0000	0.0000	0.0000 0.0000	.	.	.
insulin	1	-0.1327	0.4109	-0.9380 0.6726	0.10	0.10	0.7467
duration	1	0.0034	0.0022	-0.0010 0.0078	2.30	2.30	0.1291
female	1	0.0611	0.1685	-0.2691 0.3913	0.13	0.13	0.7168
adult	1	0.7487	0.3387	0.0848 1.4125	4.89	4.89	0.0271
bcval5	1	3.7349	1.3103	1.1668 6.3030	8.12	8.12	0.0044
hbael	1	0.2033	0.0584	0.0888 0.3177	12.11	12.11	0.0005
hxcoma	1	-1.2231	0.4933	-2.1900 -0.2563	6.15	6.15	0.0132

Within the i th matched set, the conditional probability of $\{d_{i1}, \dots, d_{in_i}\}$ events among the n_i members with covariate vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_{n_i}\}$, respectively, is

$$P[\{(d_{i1}, \mathbf{x}_1), \dots, (d_{in_i}, \mathbf{x}_{n_i})\} | d_{i\bullet}, n_i]. \quad (8.59)$$

It is then readily shown that the conditional likelihood is

$$L_{(c)}(\boldsymbol{\beta}) = \prod_{i=1}^N L_{i|d_{i\bullet}, n_i} = \prod_{i=1}^N \frac{\prod_{j=1}^{n_i} \left[e^{\mathbf{x}'_{ij} \boldsymbol{\beta} + \log(t_{ij})} \right]^{d_{ij}}}{\sum_{\ell=1}^{n_i!} \prod_{j(\ell)=1}^{n_i} \left[e^{\mathbf{x}'_{ij(\ell)} \boldsymbol{\beta} + \log(t_{ij(\ell)})} \right]^{d_{ij}}}, \quad (8.60)$$

(see Problem 8.8). Standard software is not available to fit this model; however, it could be fit using the general PROC NLIN.

For the special case of matched pairs, this conditional likelihood reduces to

$$\begin{aligned} L_{(c)}(\boldsymbol{\beta}) &= \prod_{i=1}^N \left(1 + \left[e^{(\mathbf{x}_{i1} - \mathbf{x}_{i2}) \boldsymbol{\beta}} \left(\frac{t_{i1}}{t_{i2}} \right) \right]^{d_{i2} - d_{i1}} \right)^{-1} \\ &= \prod_{i=1}^N (1 + e^{-\Delta\eta})^{-1}, \end{aligned} \quad (8.61)$$

where

$$\Delta\eta = (d_{i1} - d_{i2}) \left[(\mathbf{x}_{i1} - \mathbf{x}_{i2}) \boldsymbol{\beta} + \log \left(\frac{t_{i1}}{t_{i2}} \right) \right]. \quad (8.62)$$

The latter is a logistic regression model with no intercept, covariate vector $(d_{i1} - d_{i2}) (\mathbf{x}_{i1} - \mathbf{x}_{i2})$, and offset $(d_{i1} - d_{i2}) \log(t_{i1}/t_{i2})$; and where the binary dependent variable $y_i = 1$ for all pairs.

In either case, the score equations and information matrix are readily obtained.

8.6 NEGATIVE BINOMIAL MODELS

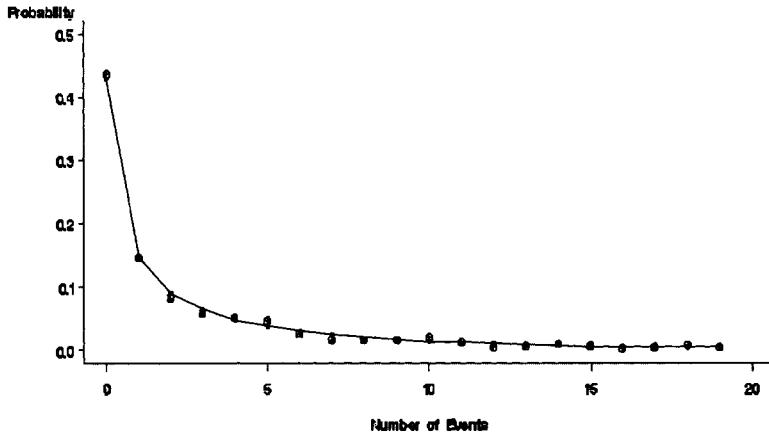
8.6.1 The Negative Binomial Distribution

As suggested by Figure 8.1, neither the Poisson nor the zeros-inflated Poisson models provide a good fit to the distribution of numbers of episodes of hypoglycemia in the DCCT data. A natural alternative in such cases is the negative binomial distribution with a rate parameter λ and a dispersion parameter κ , where the probability distribution of the number of events over exposure time t is of the form

$$\Pr(d; \lambda, \kappa, t) = \frac{\Gamma(d + 1/\kappa)(\lambda t \kappa)^d}{\Gamma(d + 1)\Gamma(1/\kappa)(1 + \lambda t \kappa)^{d+1/\kappa}}, \quad d = 0, 1, 2, \dots \quad (8.63)$$

For given t , then $E(d) = \lambda t$, and $V(d) = \lambda t + \kappa(\lambda t)^2$. The parameter κ is termed the negative binomial dispersion parameter and is one of the natural parameters in

Fig. 8.2 Negative binomial model fit to the overall DCCT hypoglycemia data. The empirical distribution is designated by \oplus .



the model. It is not the same as the general overdispersion scale parameter used previously to allow for overdispersion in a quasi-likelihood regression model.

The corresponding log likelihood is

$$\ell(\lambda, \kappa) = d \log(\lambda t \kappa) - (d + 1/\kappa) \log(1 + \lambda t \kappa) + \log \left[\frac{\Gamma(d + 1/\kappa)}{\Gamma(d + 1) \Gamma(1/\kappa)} \right] \quad (8.64)$$

$$\propto d \log \left[\frac{\lambda t \kappa}{1 + \lambda t \kappa} \right] - (1/\kappa) \log(1 + \lambda t \kappa),$$

that is the form of a generalized linear model as in Section A.10 with the log as the canonical link [see Lawless (1987) and McCullagh and Nelder, (1989)]. As $\kappa \rightarrow 0$, the negative binomial log likelihood approaches that of the Poisson.

Example 8.9 Hypoglycemia in the DCCT (continued)

First consider a negative binomial model fit to the complete cohort with no covariates. The SAS statements are

```
proc genmod;
model nevents = / dist = negbin link = log offset = lnyears;
```

As with the Poisson model, the offset allows for varying degrees of exposure in the cohort. The resulting model fit statistics are

Criteria For Assessing Goodness Of Fit				
Criterion	DF	Value	Value/DF	
Deviance	714	702.7525	0.9842	
Pearson Chi-Square	714	719.1964	1.0073	

The value/df is close to 1 for both measures indicating a good fit. The parameter estimates are

Analysis Of Maximum Likelihood Parameter Estimates						
	Parameter	DF	Standard Estimate	Wald 95% Confidence Limits	Wald ChiSq	Pr > ChiSq
	Intercept	1	-0.8361	0.0641 -0.9617	-0.7104	170.09 <.0001
	Dispersion	1	2.6059	0.1867 2.2400	2.9719	.

The intercept is the $\log(\lambda)$ so that $\hat{\lambda} = e^{-0.8361} = 0.4334$ per year or 43.3 per 100 patient years. This is close to the overall crude rate of 44.6 per 100 patient years computed in Example 8.1.

Evaluating the model in (8.63) at an average follow-up time of $t = 7.1$ years yields the distribution shown in Figure 8.2. It is clear that the negative binomial distribution provides nearly perfect fit to the overall observed data.

8.6.2 Negative Binomial Regression Model

As in Section 8.3.1, we now assume that the rate parameter $\lambda(\mathbf{x}_i)$ for the i th subject can be expressed as a log-linear function of a vector of covariates \mathbf{x}_i with an intercept (α), coefficient vector β as in (8.45), and that the expected count $E(d_i|\mathbf{x}_i, t_i) = \lambda(\mathbf{x}_i)t_i$ can be expressed as in (8.46) with $offset = \log(t_i)$. Then the model likelihood is the product of terms like (8.63), substituting the values d_i and $\lambda(\mathbf{x}_i)t_i$ for the i th subject. Maximum likelihood estimation then provides estimates of the parameters $\boldsymbol{\theta} = (\alpha, \kappa, \beta)$.

For the case of matched sets, the unconditional model would employ set-specific values $\boldsymbol{\theta}_i = (\alpha_i, \kappa_i, \beta_i)$ for the i th such set. From the log likelihood in (8.64), it is clear that conditioning on the total numbers of events within each set does not eliminate the nuisance parameters (α_i, κ_i) from the conditional likelihood.

Example 8.10 Hypoglycemia in the DCCT (continued)

First consider the assessment of the overall treatment group effect estimate that is provided by this model using the statements

```
proc genmod;
  model nevents = / dist = negbin link = log offset = lnyears;
  lsmeans conv / cl;
```

This provides the model estimates as follows:

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Standard Estimate	Error	Wald Confidence Limits	ChiSq	Wald ChiSq	Pr >
Intercept	1	-1.5510	0.0902	-1.7278 -1.3742	295.60	<0.0001	
conv	0	1	1.1173	0.1215 0.8791	1.3555	84.52	<0.0001
conv	1	0	0.0000	0.0000 0.0000	0.0000	.	
Dispersion	1	2.1863	0.1649	1.8631	2.5095		

The model is highly significant with a deviance of 708.0 on 713 *df* and a value/*df* = 0.9930.

The group coefficient estimate (intensive vs. conventional) provides an estimated relative risk of $\widehat{RR} = \exp(1.1173) = 3.057$ with 95% confidence limits $\exp(0.8791, 1.3555) = (2.41, 3.88)$. The point estimate is similar to that provided by the various Poisson models, but the standard error of the estimate (0.1215), or of the $\log(\widehat{RR})$, is substantially greater than that estimated by the homogeneous Poisson model in Table 8.4 (0.0492), or that obtained from a ZIP model (0.0518, not previously shown). The unadjusted rate, per 100 patient years, in the conventional group is $100 \exp(-1.5510) = 21.20$ versus that in the intensive group of $100 \exp(-1.551 + 1.1173) = 64.811$. These estimates are also provided by the *lsmeans* statement.

A model with treatment group adjusted for other factors would be fit using the statements

```
proc genmod; class conv;
  model nevents = conv insulin duration female
    adult bcvals hbael hxcoma
  / dist = NegBin link = log offset = lnyears;
  lsmeans conv / cl;
```

This yields the results shown in Table 8.12. Compared to the homogeneous Poisson model in Table 8.5, with the exception of *hxcoma*, all of the coefficients have shifted away from zero, some dramatically, such as *bcvals*, where the coefficient changed from -0.53 to -1.63. The standard errors are also much larger and are similar to those provided by the quasi-likelihood Poisson model with a fixed overdispersion parameter (Table 8.7), and the GEE model with the empirical robust covariance matrix estimate (Table 8.9), both of which also employ the same coefficients as in Table 8.5.

The adjusted relative risk for treatment group is $\widehat{RR} = \exp(1.1996) = 3.319$, that is larger than estimated from the Poisson models (2.96). The adjusted rate within each group provided by the *lsmeans* statement, per 100 patient years, is 63.4 for the intensive group versus 19.1 for the conventional group.

Table 8.12 PROC GENMOD negative binomial regression analysis of the DCCT data.

The GENMOD Procedure							
Criteria For Assessing Goodness Of Fit							
Criterion	DF	Value		Value/DF			
Deviance	706	707.6895		1.0024			
Pearson Chi-Square	706	718.1531		1.0172			

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Error	Standard	Wald 95%	Wald	Pr >
Intercept	1	-0.5087	0.6432	-1.7694	0.7520	0.63	0.4290
conv	0	1.1996	0.1231	0.9583	1.4410	94.89	<.0001
conv	1	0.0000	0.0000	0.0000	0.0000	.	.
insulin	1	0.2136	0.2881	-0.3511	0.7782	0.55	0.4585
duration	1	0.0000	0.0017	-0.0033	0.0034	0.00	0.9840
female	1	0.2471	0.1231	0.0057	0.4884	4.03	0.0448
adult	1	-0.6541	0.2161	-1.0776	-0.2305	9.16	0.0025
bcval5	1	-1.6283	1.0058	-3.5997	0.3430	2.62	0.1055
hbael	1	-0.0817	0.0424	-0.1648	0.0015	3.70	0.0543
hxcoma	1	0.4798	0.2582	-0.0263	0.9859	3.45	0.0631
Dispersion	1	2.0299	0.1560	1.7242	2.3356		

Least Squares Means							
Effect	Mean	Estimate	Standard	95%	Confidence	Limits	
conv	0	0.6340	-0.4557	0.0804	-0.6133	-0.2982	
conv	1	0.1910	-1.6554	0.0909	-1.8335	-1.4772	

8.7 POWER AND SAMPLE SIZE

8.7.1 Poisson Models

First consider the power of the simple Z -test for two groups in (8.20) of $H_0: \lambda_1 = \lambda_2 = \lambda$ versus $H_1: \lambda_1 \neq \lambda_2$ under a homogeneous Poisson model. As in Chapter 3, denote the sample fractions within each group as ξ_i , where $n_i = N\xi_i$ ($i = 1, 2$). For evaluation of sample size or power for given λ_1 and λ_2 , it is generally assumed under H_0 that $\lambda = \xi_1\lambda_1 + \xi_2\lambda_2$. In Problem 3.5 we derived the expressions for the power of the test assuming a common unit of exposure for all observations. In Problem 8.7, an alternative expression is presented as a function of the total expected numbers of events and total exposure time.

Assume that the distribution of the exposure times $\{t_{ij}\}$ is the same among the subjects in the two groups with mean η_t . Then $E(t_{i\bullet}) = N\xi_i\eta_t$. From (8.88) of

Problem 8.7, it follows that the basic relationship is

$$|\lambda_1 - \lambda_2| \sqrt{N} = Z_{1-\alpha} \sqrt{\frac{\lambda}{\xi_1 \xi_2 \eta_t}} + Z_{1-\beta} \sqrt{\frac{\lambda_1}{\xi_1 \eta_t} + \frac{\lambda_2}{\xi_2 \eta_t}}, \quad (8.65)$$

that yields an expression for $Z_{1-\beta}$ as a function of the other parameters for a given N , or the expression for N .

If extra-variation is observed or is anticipated, then a simple approach would be to adopt the overdispersed quasi-likelihood model with a scale or dispersion factor ν , where $V(d_i) = \nu \lambda t_i$. Then the above expressions would be modified to simply employ $\nu \lambda_1$, $\nu \lambda_2$, and $\nu \lambda$.

Alternatively, the power function of the Z -test in (8.44) in the two-stage random effects model can be employed. Let μ_{λ_i} denote the mixing distribution mean within the i th group under the alternative hypothesis $H_1: \mu_{\lambda_1} \neq \mu_{\lambda_2}$; and let μ_λ denote the common mean rate within both groups under the null hypothesis $H_0: \mu_{\lambda_1} = \mu_{\lambda_2} = \mu_\lambda$. Then let $\sigma_{\lambda_i(0)}^2$ denote the overdispersion variance component within the i th group under the null hypothesis that is the expected value of (8.37) with respect to μ_λ . Likewise, let $\sigma_{\lambda_i(1)}^2$ denote the variance within each group evaluated under the alternative. Within the i th group, under hypothesis H_ℓ ($\ell = 0, 1$), the variance of the mean rate (\bar{r}_i) for given N is obtained as the expectation of (8.38) with respect to the distribution of the exposure times t_{ij} among the subjects within each group. Assume that the distribution of exposure times is the same in the two groups with mean η_t and variance v_t^2 . Then

$$E \left(\sum_{j=1}^{n_i} t_{ij}^2 \right) = N \xi_i E(t_{ij}^2) = N \xi_i (v_t^2 + \eta_t^2) \quad (8.66)$$

and

$$V(\bar{r}_i | H_\ell) = \frac{\sigma_{\lambda_i(\ell)}^2 [v_t^2 + \eta_t^2] + \mu_{\lambda_i(\ell)} \eta_t}{N \xi_i \eta_t^2} = \frac{\phi_{i(\ell)}^2}{N}. \quad (8.67)$$

Then, the basic relationship is

$$|\mu_{\lambda_1} - \mu_{\lambda_2}| \sqrt{N} = Z_{1-\alpha} \left[\phi_{1(0)}^2 + \phi_{2(0)}^2 \right]^{1/2} + Z_{1-\beta} \left[\phi_{1(1)}^2 + \phi_{2(1)}^2 \right]^{1/2},$$

that is readily solved for $Z_{1-\beta}$ or N . If it is assumed that $t_{ij} = \eta_t$ for all observations (i.e., a fixed exposure time), then

$$\phi_{i(\ell)}^2 = \frac{\sigma_{\lambda_i(\ell)}^2 \eta_t + \mu_{i(\ell)}}{\xi_i \eta_t}. \quad (8.68)$$

In the Poisson regression model, the power of the Wald test with one or more qualitative covariates is also readily obtained along the lines of that for the Wald test in logistic regression, as described in Section 7.3.6. The resulting equations are also described in Problem 8.7.

The power of a Wald test in an overdispersed Poisson regression model may also be obtained using $V(\hat{\theta}) = \nu \Sigma_\theta$, where ν is the specified overdispersion parameter

and Σ_{θ} is the model-based covariance matrix of the estimates. Alternatively, for power computations for a test based on an observed data set, the robust estimate of the covariance matrix of the estimates may be used.

8.7.2 Negative Binomial Models

As for the Poisson model, we start with the properties of a simple Z -test for the difference between groups for a known or specified value of the dispersion parameter κ . This test would not be used in practice; rather, a regression model would best be employed that provides joint estimation of the treatment group coefficient and κ . However, for a given value of κ , this test can provide the basis for the evaluation of sample size or power for the model-based test of the group effect in the model.

From (8.64) the log likelihood for a sample of N observations (d_j, t_j) is

$$\ell(\lambda, \kappa) = \sum_{j=1}^N d_j \log [\lambda t_j \kappa] - \sum_{j=1}^N d_j \log [1 + \lambda t_j \kappa] - \sum_{j=1}^N (1/\kappa) \log(1 + \lambda t_j \kappa). \quad (8.69)$$

Assuming that the dispersion parameter κ is known or specified, the score or estimating equation for λ is

$$U(\lambda) = \sum_{j=1}^N \left[\frac{d_j}{\lambda} - \frac{d_j t_j \kappa}{1 + \lambda t_j \kappa} - \frac{t_j}{1 + \lambda t_j \kappa} \right], \quad (8.70)$$

that when set equal to zero yields

$$U(\lambda) = 0 = \sum_{j=1}^N [d_j - \lambda t_j], \quad (8.71)$$

so that

$$\hat{\lambda} = \frac{\sum_{j=1}^N d_j}{\sum_{j=1}^N t_j} = \frac{d_{\bullet}}{t_{\bullet}}, \quad (8.72)$$

as for the Poisson distribution. Since $V(d_j) = \lambda t_j + \kappa(\lambda t_j)^2$, then it follows that

$$V(\hat{\lambda}) = \frac{\sum_{j=1}^N [\lambda t_j + \kappa(\lambda t_j)^2]}{\left(\sum_{j=1}^N t_j\right)^2} = \frac{\lambda t_{\bullet} + \lambda^2 \kappa \left(\sum_{j=1}^N t_j^2\right)}{t_{\bullet}^2}. \quad (8.73)$$

Applying the delta method, then

$$V(\log \hat{\lambda}) = \frac{V(\hat{\lambda})}{\lambda^2} = \frac{t_{\bullet} + \lambda \kappa \left(\sum_{j=1}^N t_j^2\right)}{\lambda t_{\bullet}^2}. \quad (8.74)$$

For the comparison of two groups with sample sizes n_1 and n_2 , let $\hat{\theta} = \log(\widehat{RR}) = \log(\widehat{\lambda}_1) - \log(\widehat{\lambda}_2)$ with estimated variance $\widehat{V}(\hat{\theta}) = \widehat{V}(\log \widehat{\lambda}_1) + \widehat{V}(\log \widehat{\lambda}_2)$, where

$\widehat{V}(\log \widehat{\lambda}_i)$ is obtained from the above with n_i substituted for N and $\widehat{\lambda}$ for λ and with fixed κ . Then a Wald test is provided by $Z = \widehat{\theta}/\sqrt{\widehat{V}(\widehat{\theta})}$. This Z -test would not be used in practice because it requires that κ be known or specified. However, for a value κ derived from a regression model such as that presented above, the resulting estimates and Z -value will be close to the model estimated values.

For the DCCT data presented in Example 8.10, the above computations yield $\widehat{\theta} = 1.1080$ with estimated variance $\widehat{V}(\widehat{\theta}) = 0.01524$, $S.E. = 0.1235$, and $Z = 8.98$, or $\chi^2 = 80.57$, all close to the values provided by the negative binomial model with a group effect. Thus, the power of the model test of the group effect, when also estimating the dispersion parameter κ , can be approximated by the power of the above Z -test for given values $(\lambda_1, \lambda_2, \kappa)$ with an assumed distribution of exposure times.

Under $H_0: \lambda_1 = \lambda_2 = \lambda$, then

$$V(\widehat{\theta}|H_0) = \sum_{i=1}^2 \frac{t_{i\bullet} + \lambda\kappa \left(\sum_{j=1}^{n_i} t_{ij}^2 \right)}{\lambda t_{i\bullet}^2}, \quad (8.75)$$

and under $H_1: \lambda_1 \neq \lambda_2$,

$$V(\widehat{\theta}|H_1) = \sum_{i=1}^2 \frac{t_{i\bullet} + \lambda_i\kappa \left(\sum_{j=1}^{n_i} t_{ij}^2 \right)}{\lambda_i t_{i\bullet}^2}. \quad (8.76)$$

Each is evaluated with respect to the distribution of the exposure times $\{t_{ij}\}$ among the subjects within each group. Again assume a common distribution in the two groups with mean η_t and variance v_t^2 and sample fractions (ξ_1, ξ_2) within each group. Then

$$\begin{aligned} V(\widehat{\theta}|H_0) &= \sum_{i=1}^2 \frac{\eta_t + \lambda\kappa [v_t^2 + \eta_t^2]}{\lambda N \xi_i \eta_t^2} = \frac{\phi_{(0)}^2}{N} \\ V(\widehat{\theta}|H_1) &= \sum_{i=1}^2 \frac{\eta_t + \lambda_i\kappa [v_t^2 + \eta_t^2]}{\lambda_i N \xi_i \eta_t^2} = \frac{\phi_{(1)}^2}{N} \end{aligned} \quad (8.77)$$

and the basic relationship is

$$|\log(\lambda_1/\lambda_2)|\sqrt{N} = Z_{1-\alpha}\phi_{(0)} + Z_{1-\beta}\phi_{(1)} \quad (8.78)$$

that is readily solved for $Z_{1-\beta}$ or N .

Illustrative computations are presented in Problems 8.7 and 8.12.

8.8 MULTIPLE OUTCOMES

Chapter 7 describes the application of generalized estimation equations to the analysis of repeated binary measurements over time, or multiple separate (multivariate) measurements. Those techniques can also be applied to the analysis of repeated

or multivariate count data measurements. For example, Thall and Vail (1990) describe the analysis of a clinical trial in which the numbers of epileptic seizures were recorded at four successive visits at two-week intervals, where some subjects could have missed a visit in which case the corresponding count is missing. Treating these as repeated count data allows the conduct of an overall inference with respect to the effects of treatment on the incidence of seizures over the full eight-week period. This analysis, and the study data, are provided in the *SAS/STAT User's Guide* (2008b).

8.9 PROBLEMS

8.1 Consider the doubly homogeneous Poisson process described in Section 8.1.2.

8.1.1. Show that the crude rate in (8.9) under a doubly homogeneous Poisson model can be expressed a weighted least squares (inverse variance weighted) estimate of the form

$$\bar{r} = \frac{\sum_j \hat{\tau}_j r_j}{\sum_j \hat{\tau}_j}, \quad (8.79)$$

where $\hat{\tau}_j = \hat{V}(r_j)^{-1}$.

8.1.2. Then show that its estimated variance equals $\hat{V}(\bar{r}) = \left[\sum_j \hat{\tau}_j \right]^{-1} = \hat{V}(\hat{\lambda})$ in (8.10).

8.2 Now consider the robust information sandwich estimate of the variance as described in Section A.9.

8.2.1. Based on the score equation in (8.7), show that the empirical estimate of the observed information is

$$\hat{J}(\hat{\lambda}) = \frac{\sum_{j=1}^N (d_j - \hat{\lambda} t_j)^2}{\hat{\lambda}^2}. \quad (8.80)$$

8.2.2. Also show that the robust estimate of the variance of the estimate is

$$\hat{V}_R(\hat{\lambda}) = \frac{\sum_{j=1}^N (d_j - \hat{\lambda} t_j)^2}{t_\bullet^2}. \quad (8.81)$$

8.2.3. Under the doubly homogeneous assumptions show that

$$E \left[\hat{V}_R(\hat{\lambda}) \right] = \frac{\lambda}{t_\bullet} = I(\lambda)^{-1}. \quad (8.82)$$

8.3 Consider a random effects overdispersed model as in Section 8.2.

8.3.1. To facilitate computation of the variance component, show that the moment estimator of the dispersion variance component presented in (8.37) can also be

expressed as

$$\hat{\sigma}_{\lambda}^2 = \max \left[0, \frac{\sum_j \frac{(d_j - \hat{\mu}_{\lambda} t_j)^2}{t_j} - n \hat{\mu}_{\lambda}}{\sum_j t_j} \right]. \quad (8.83)$$

8.3.2. Also show that this can be expressed as a weighted mean of variables a_j of the form

$$\hat{\sigma}_{\lambda}^2 = \frac{\sum_j t_j a_j}{\sum_j t_j}, \quad a_j = \frac{1}{t_j} \left[\frac{(d_j - \hat{\mu}_{\lambda} t_j)^2}{t_j} - \hat{\mu}_{\lambda} \right]. \quad (8.84)$$

8.4 Gail et al. (1980) present data on the numbers of tumor recurrences over a period of 122 days among 23 female rats treated with retinoid versus 25 untreated controls. The following are the numbers of tumor recurrences for the rats in each group (reproduced with permission):

Retinoid: 1, 0, 2, 1, 4, 3, 6, 1, 1, 5, 2, 1, 5, 2, 3, 4, 5, 5, 1, 2, 6, 0, 1

Control: 7, 11, 9, 2, 9, 4, 6, 7, 6, 1, 13, 2, 1, 10, 4, 5, 11, 11, 9, 12,

1, 3, 1, 3, 3

8.4.1. Assuming a doubly homogeneous model, compute the estimate of the Poisson rate (per day) for tumor recurrence in each group ($\hat{\lambda}_1, \hat{\lambda}_2$) and in the combined sample of 48 rats ($\hat{\lambda}$).

8.4.2. Also compute the large sample variance of the log rate and the asymmetric 95% confidence limits for the rate in each group and in the combined sample.

8.4.3. Compute the relative risk and the asymmetric 95% confidence limits based on the estimated variance of the log(\widehat{RR}).

8.4.4. Compute the estimated variance of the difference in rates under the null hypothesis $\widehat{V}(\hat{\lambda}_1 - \hat{\lambda}_2 | H_0)$ and the simple Z -test of the difference in rates between groups.

8.4.5. Compute Cochran's variance test of the hypothesis of no overdispersion separately within each treatment group.

8.4.6. Using (8.81), compute the information sandwich robust estimate of the variance of the estimated rate within each group $\widehat{V}_R(\hat{\lambda}_i)$, $i = 1, 2$.

8.4.7. Use this with (8.15) to compute the robust estimate of the variance of $\log(\hat{\lambda}_1)$ and $\log(\hat{\lambda}_2)$.

8.4.8. Then compute the robust estimate of the variance of $\log(\widehat{RR})$ and the robust 95% confidence limits for RR .

8.4.9. Alternatively, adopt a two-stage model as in Section 8.2. Using (8.37), compute the moment estimate of the overdispersion variance $\hat{\sigma}_{\lambda_i}^2$ for each group ($i = 1, 2$).

8.4.10. Use this to compute an overdispersion estimate of the variance of the crude rate within each group as in (8.38) and of the log rate as in (8.39). Use these to

compute the overdispersed estimate of the variance of the $\log(\widehat{RR})$. Compare these to those obtained using the robust estimate of the variance.

8.5 Now consider the case of two independent groups where the observations are divided into K independent strata.

8.5.1. Assume that the rates in the control group, $\{\lambda_{2j}\}$, vary among strata but that there is a common log relative risk among the K strata, $\theta_j = \log(\lambda_{1j}/\lambda_{2j}) = \theta$, $j = 1, \dots, K$. Let $d_{1\bullet j}$ and $t_{1\bullet j}$ denote the total numbers of events and total period of exposure, respectively, in the i th group within the j th stratum. Show that the *MVLE* of the common $\log(RR)$ is provided by

$$\widehat{\theta} = \frac{\sum_{j=1}^K \left(\frac{d_{1\bullet j} d_{2\bullet j}}{d_{1\bullet j} + d_{2\bullet j}} \right) \log(\widehat{\lambda}_{1j}/\widehat{\lambda}_{2j})}{\sum_{\ell=1}^K \left(\frac{d_{1\bullet \ell} d_{2\bullet \ell}}{d_{1\bullet \ell} + d_{2\bullet \ell}} \right)} \quad (8.85)$$

and its large sample variance is estimated as

$$\widehat{V}(\widehat{\theta}) = \left[\sum_{\ell=1}^K \left(\frac{d_{1\bullet \ell} d_{2\bullet \ell}}{d_{1\bullet \ell} + d_{2\bullet \ell}} \right) \right]^{-1}. \quad (8.86)$$

8.5.2. Using the developments in Section 4.7.2, derive the expression for the maximally efficient asymptotic test of the joint null hypothesis $H_0: \theta_j = 0$ for all K strata against the alternative that a common log relative risk applies over all strata $H_1: \theta_j = \theta \neq 0$. Show that the test is of the form

$$X_A^2 = \frac{\left[\sum_{j=1}^K \left(\frac{t_{1\bullet j} t_{2\bullet j}}{t_{1\bullet j} + t_{2\bullet j}} \right) (\widehat{\lambda}_{1j} - \widehat{\lambda}_{2j}) \right]^2}{\sum_{j=1}^K \left(\frac{t_{1\bullet j} t_{2\bullet j}}{t_{1\bullet j} + t_{2\bullet j}} \right) \widehat{\lambda}_j}. \quad (8.87)$$

8.5.3. Show that Cochran's test of the hypothesis of homogeneity of relative risks among strata, analogous to the Cochran test for independent 2×2 tables in Section 4.10.2, is given by

$$X_{H,C}^2 = \sum_{j=1}^K \widehat{\tau}_j (\widehat{\theta}_j - \widehat{\theta})^2,$$

where $\widehat{\tau}_j = d_{1\bullet j} d_{2\bullet j} / (d_{1\bullet j} + d_{2\bullet j})$, and where $X_{H,C}^2$ is distributed as chi-square on $K-1$ df under the hypothesis of homogeneity $H_0: \theta_j = \theta \forall j$.

8.5.4. Now adopt a two-stage random effects model across strata. The random effects model estimate of the variance component between strata $\widehat{\sigma}_\theta^2$ is obtained directly from (4.154) using the above expression for $\widehat{\tau}_j$. From this, derive the resulting updated (one-step) estimate of the mean $\log(RR)$ over strata and its variance.

8.6 Table 8.13 presents data from Frome and Checkoway (1985) giving the numbers of cases of nonmelanoma skin cancer and the population size within the Dallas/Fort Worth versus Minneapolis/St. Paul metropolitan areas, stratified by age groups. The

Table 8.13 Numbers of cases of skin cancer and population size, stratified by age, in Dallas/Fort Worth and Minneapolis/St. Paul, from Frome and Checkoway (1985), reproduced with permission.

Age	Dallas/Fort Worth		Minneapolis/St. Paul		\widehat{RR}
Stratum	d_1	T_1	d_2	T_2	
15–24	4	181,343	1	172,675	3.81
25–34	38	146,207	16	123,065	2.00
35–44	119	121,374	30	96,216	3.14
45–54	221	111,353	71	92,051	2.57
55–64	259	83,004	102	72,159	2.21
65–74	310	55,932	130	54,722	2.33
75–84	226	29,007	133	32,185	1.89
85+	65	7,538	40	8,328	1.80
Total	1242	735758	523	651401	2.10

hypothesis is that those in the Dallas area would have a higher rate of such cancer because of greater exposure to the sun.

8.6.1. Based on the aggregate data for all age strata combined, compute the unadjusted relative risk for location (Dallas vs. Minneapolis), the $\log(\widehat{RR})$, its estimated variance and the asymmetric 95% confidence limits for the RR .

8.6.2. Compute the unadjusted Z -test of the hypothesis of equal rates in the two metropolitan areas.

8.6.3. Compute the *MVLE* of the stratified-adjusted common log relative risk, its variance, and the asymmetric confidence limits for the common relative risk.

8.6.4. Compute the stratified-adjusted efficient test of a common relative risk and compare the results to the unadjusted test.

8.6.5. Compute Cochran's test of homogeneity and the estimate of the random effects model variance component.

8.6.6. Then compute the one-step estimate of the mean log relative risk and its variance under the random effects model. Also compute the asymmetric confidence limits for the RR . Compare these to the *MVLE* estimate and its limits.

8.6.7. Alternatively, these data could be analyzed using Poisson regression. Fit a Poisson regression model with an effect for location and binary indicator variables for seven of the eight strata. Compare the model-based estimate of the relative risk for location to the *MVLE* above.

8.7 Show that the power function of the Z -test for the difference in rates between two groups in (8.26) under the doubly homogeneous Poisson model is a function of

the total patient years of exposure of the form

$$\begin{aligned} |\lambda_1 - \lambda_2| &= Z_{1-\alpha} \sqrt{\lambda \left(\frac{t_{1\bullet} + t_{2\bullet}}{t_{1\bullet} t_{2\bullet}} \right)} + Z_{1-\beta} \sqrt{\frac{\lambda_1}{t_{1\bullet}} + \frac{\lambda_2}{t_{2\bullet}}} \\ &= Z_{1-\alpha} \sqrt{\lambda^2 \left(\frac{E(d_{1\bullet} + d_{2\bullet})}{E(d_{1\bullet})E(d_{2\bullet})} \right)} + Z_{1-\beta} \sqrt{\frac{\lambda_1^2}{E(d_{1\bullet})} + \frac{\lambda_2^2}{E(d_{2\bullet})}}, \end{aligned} \quad (8.88)$$

which is also a function of the expected number of events in each group.

8.7.1. Assume that we wished to detect a difference between groups of $\lambda_1 = 0.25/\text{year}$ versus $\lambda_2 = 0.15/\text{year}$ with two equal-sized patient groups, that is, $t_{1\bullet} = t_{2\bullet} = (t_{1\bullet} + t_{2\bullet})/2$. At $\alpha = 0.05$ two-sided:

1. Show that a total of 840.6 patient years are required to provide 90% power to detect this difference assuming a doubly homogeneous model. If the mean exposure time is $\eta_t = 3$ years then this is equivalent to a total $N = 280.2$ (or 282). Conversely, if the total $N = 400$, then a mean follow-up of 2.1 years is required.

2. Assume that an overdispersed model applies with variance component $\sigma_\lambda^2 = 0.06$ within each of the two groups under H_0 and H_1 , and that the mean exposure time is $\eta_t = 3$ with variance $v_2^2 = 0.25$. Show that a total $N = 539.4$ (540) is required to provide 90% power to detect the same difference in the mean rates.

8.7.2. Now consider a Poisson regression model with grouped or discrete covariates analogous to the logistic regression model with discrete covariates. In this case, the developments in Section 7.3.6 can be used to derive the sample size for, or power of, a Wald test of coefficient effects. Let t_\bullet denote the total exposure time in the study that is a function of the total sample size N . Then denote the expected fraction of total exposure time in the i th cell as $\zeta_i = E(t_i/t_\bullet)$, where $t_\bullet = \sum_i t_i$, and show that $\Omega = \text{diag}\{[\zeta_i \lambda(\mathbf{x}_i)]^{-1}\}$, where the expected number of events in the i th cell with covariate vector \mathbf{x}_i is $E(d_i) = \zeta_i t_\bullet \lambda(\mathbf{x}_i)$. The remainder of the developments in Section 7.3.6 then apply.

8.7.3. Consider a model to assess the effect of a treatment or exposure group within three strata with a design matrix as presented in Example 7.9. Assume that the fractions of patient years of exposure within each of the six cells are $\{\zeta_i\} = (0.08, 0.12, 0.30, 0.20, 0.125, 0.175)$. The assumed intensities per year in each cell are $\{\lambda_i\} = (0.15, 0.28, 0.12, 0.27, 0.20, 0.42)$. Generate a large data set with event counts within the i th cell proportional to the Poisson probability (λ_i) times the cell sample fraction (ζ_i). Then fit a Poisson regression model with binary indicator variables for stratum 2 versus 1 (x_1), stratum 3 versus 1 (x_2), and treatment group (x_3). The coefficients (θ) are approximately $\alpha = -1.9929$, $\beta_1 = -0.0925$, $\beta_2 = 0.3759$, and $\beta_3 = 0.7520$, the latter corresponding to a relative risk of 2.12 comparing the treatment groups.

8.7.4. Use the values of $\{\mathbf{x}_i \theta\}$ and $\{\zeta_i\}$ to compute Ω and substitute this into (7.77) to obtain the covariance matrix of the estimates. Then, determine the value of the noncentrality factor τ^2 for the 1 df test of the adjusted group log relative risk, $H_0: \beta_3 = 0$, with vector $\mathbf{C}' = (0 \ 0 \ 0 \ 1)$. Then, show that the total patient years required to provide 90% power for this test is $t_\bullet = 320.7$.

8.7.5. Conversely, for the same model, show that a total exposure of only $t_{\bullet} = 200$ years provides power of 0.651 to detect an adjusted group relative risk of 2.12.

8.8 Consider a Poisson regression model with a single binary covariate for treatment group ($x = 1 = E, 0 = \bar{E}$).

8.8.1. Assuming that $\beta = 0$, show that $U(\alpha) = 0$ implies that

$$e^{\alpha} = \frac{\sum_i d_i}{\sum_i t_i} = \frac{d_{\bullet}}{t_{\bullet}}. \quad (8.89)$$

8.8.2. Derive the elements of $I(\hat{\theta}_0)$.

8.8.3. Show that the score test for β equals the Z -test presented in (8.25).

8.9 Derive the expressions for the proportion of explained variation in the Poisson regression model.

8.9.1. $\rho_{\varepsilon^2}^2$ assuming squared error loss.

8.9.2. R_{resid}^2 and show that this equals $\hat{\rho}_{\varepsilon^2}^2$ in (8.52).

8.10 Fleming and Harrington (1991) present the results of a randomized clinical trial of the effects of gamma interferon versus placebo on the incidence of serious infections among children with chronic granulomatous disease (CGD). For each subject the number of infections experienced and the total duration of follow-up are presented. In Chapter 9 an analysis is presented using the actual event times. Here we consider only the number of events experienced by each subject. The count data set is available online (see the Preface, reproduced with permission). The data set includes the patient *id*, number of severe infections experienced (*nevents*), the number of days of follow-up (*futime*), and the following covariates: Z_1 : treatment group: interferon (1) versus placebo (0); Z_2 : inheritance pattern: X-linked (1) versus autosomal recessive (2); Z_3 : age (years); Z_4 : height (cm); Z_5 : weight (kg); Z_6 : corticosteroid use on entry: yes (1) versus no (2); Z_7 : antibiotic use on entry: yes (1) versus no (2); Z_8 : gender: male (1) versus female (2); and Z_9 : type of hospital: NIH (1), other US (2), Amsterdam (3), other European (4). Use these data to conduct the following analyses.

8.10.1. Compute the crude rates of infection per day in each treatment group and the asymmetric confidence limits on each using the $\text{log}(\text{rate})$ under the homogeneous Poisson assumptions.

8.10.2. Also compute the asymmetric confidence limits for the relative risk and the Z -test for the difference between groups under the homogeneous Poisson assumptions.

8.10.3. Within each group compute Cochran's variance test of the hypothesis of no overdispersion.

8.10.4. Within each group, adopt an overdispersed Poisson model and use the moment equation to compute an estimate of the mixing distribution variance. Use this to revise the computations for the variance and confidence limits on the rates within each group and the relative risk.

8.10.5. Compute the estimate of the overdispersion variance under the null hypothesis of equal intensities in each group and use these estimates to compute a Z -test to test the difference between groups allowing for overdispersion.

8.10.6. Now assume that the overdispersion can be accounted for by inclusion of covariate effects in a Poisson regression model. Assess the relative risk for interferon versus placebo in a Poisson regression model, and describe the model and the covariate effects:

1. Unadjusted.

2. Adjusted for the other covariates. Note that Z_9 should be used as a class effect to compare the four hospital types.

8.10.7. Compare the model-based results to those obtained above.

8.10.8. If the relative risks for treatment group in Problem 8.10.6.1 versus 8.10.6.2 differ, perform additional analyses as needed to help explain why.

8.10.9. Based on the analysis in Problem 8.10.3, and based on the value of Pearson X^2/df in Problem 8.10.6, is there evidence that an overdispersed model is appropriate or not?

8.10.10. Refit the model using an overdispersed quasi-likelihood model using the *pscale* option.

8.10.11. Refit the model using the information sandwich empirical estimates of the covariances using PROC GENMOD.

8.10.12. Write a program to compute the score vector and the information sandwich estimate of the observed information under the model null hypothesis. Compute the robust model score test.

8.11 Starting from the full unconditional likelihood for Poisson count data from matched pairs in (8.58):

8.11.1. Derive the score equation for the intercept α_i for the i th matched set and show that the *MLE* is a function of the sufficient statistic $d_{i\bullet} = \sum_{j=1}^{n_i} d_{ij}$.

8.11.2. Then show that the conditional likelihood from (8.59) for the i th matched set is

$$L_{i|d_{i\bullet}, n_i} = \frac{\prod_{j=1}^{n_i} \frac{e^{-\lambda(\mathbf{x}_{ij})} t_{ij}}{d_{ij}!} [\lambda(\mathbf{x}_{ij}) t_{ij}]^{d_{ij}}}{\sum_{\ell=1}^{n_i!} \prod_{j(\ell)=1}^{n_i} \frac{e^{-\lambda(\mathbf{x}_{ij(\ell)})} t_{ij(\ell)}}{d_{ij(\ell)}!} [\lambda(\mathbf{x}_{ij(\ell)}) t_{ij(\ell)}]^{d_{ij(\ell)}}}, \quad (8.90)$$

which yields (8.60).

8.11.3. For the case of matched pairs, show that (8.61) and (8.62) result.

8.11.4. For the model in (8.59), derive the elements of the score vector for β and the elements of the expected information matrix.

8.12 Now consider the negative binomial regression model described in Section 8.6.

8.12.1. Show that the negative binomial log likelihood in (8.64) approaches that of the Poisson as $\kappa \rightarrow 0$.

8.12.2. Consider the test that the log relative risk is zero for two independent groups as described in Section 8.7.2. Assume, as in Problem 8.7, that we wish to provide 90% power for a test at the 0.05 level, two-sided, with two equal-sized patient groups assuming under the alternative that $\lambda_1 = 0.25/\text{year}$ versus $\lambda_2 = 0.15/\text{year}$ ($RR = 1.67$), where $\kappa = 1.0$ and the exposure times have mean $\eta_t = 3$ years and variance $v_t^2 = 0.25$ in both groups. Show that a total sample size of $N = 442$ (rounded up) is required.

8.13 Using the data from Fleming and Harrington (1991) described in Problem 8.10, fit a no-covariate Poisson regression model and a no-covariate negative binomial model and compare the estimated probabilities of each relative to the empirical distribution as in Figures 8.1 and 8.2.

8.13.1. Based on indices of goodness of fit, which model provides a better fit to the overall data?

8.13.2. Fit a negative binomial model to compare the interferon versus placebo treatment groups without adjustment for other factors, and then with adjustment for the other covariates. Compare the group and covariate effects to those obtained from the Poisson regression models.

Analysis of Event-Time Data

Up to this point we have considered the description of the risk of an event without consideration of the *time* at which events occur in individuals in the population. One of the major advances in biostatistics has been the development of statistical methods for the analysis of event-time data, commonly known as *survival analysis*, a topic that has dominated statistical methodological research and biostatistical practice for decades. There are many excellent texts on survival analysis that cover this vast field in detail. Although originally motivated by the analysis of survival time or mortality data, these methods may be used to describe the distribution of time to any single event of interest. These methods have also been generalized to the analysis of event-times of possibly recurrent events in a subject. Many of the methods of survival or event-time analysis are generalizations of the methods developed in previous chapters.

There are many excellent general references on survival analysis. Kalbfleisch and Prentice (1980, 2002), Elandt-Johnson and Johnson (1980), Lawless (1982), and Cox and Oakes (1984), among others, give a precise review of the early methods from a classical perspective. Lee (1992), Collett (1994), and Marubini and Valsecchi (1995), among others, present thorough descriptions of the application of these methods. Fleming and Harrington (1991) and Andersen et al. (1993) present a rigorous description of the theory of martingales and counting processes and the derivation of more general methods based on these theories. Here I present a description of the principal methods for the analysis of event-time data and their application.

9.1 INTRODUCTION TO SURVIVAL ANALYSIS

9.1.1 Hazard and Survival Function

In survival or time-to-event analysis, each individual subject is followed up to some time t_i at which time either an event is observed to occur or follow-up is curtailed without observation of an event. Thus, we also observe an indicator variable δ_i that designates whether an event was observed to occur ($\delta_i = 1$) or not ($\delta_i = 0$). In the latter case, the event time is *right censored* because the actual event time is greater than the observed time of exposure. Thus, for each subject two values (t_i, δ_i) are observed that are realizations of two separate processes: the event process T and the censoring process C . Under the assumption of *censoring at random*, these two processes are assumed to be statistically independent. In this case, censored event times are considered unbiased, meaning that the same stochastic event process is assumed to apply to all observations, those censored and those not.

When the time of an event is observed to an instant of time, then event times have a right continuous, monotonically increasing distribution function $F(t) = P(T \leq t)$, $t > 0$, with a corresponding event probability density function $f(t) = dF(t)/dt$. The complement of the *cdf* is the right continuous, monotonically decreasing *survival distribution* $S(t) = 1 - F(t) = P(T > t)$. Some authors define the survival function as the left continuous $P(T \geq t)$. The distinction is irrelevant for a continuous distribution, but nontrivial when the distribution is discrete. Herein I define $S(t)$ throughout as the probability of surviving *beyond* time t so that the survival function can be defined as the complement of the cumulative distribution function for both continuous and discrete distributions.

A pivotal and informative quantity is the *hazard function* that describes the instantaneous probability of the event among those still at risk or those still free of the event. For a continuous distribution the hazard function at time t is defined as

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{P(t < T \leq t + \Delta t)}{P(T > t)} = \frac{f(t)}{S(t)} = \frac{-d \log[S(t)]}{dt}. \quad (9.1)$$

Thus, the density can also be obtained as $f(t) = \lambda(t)S(t)$. The *cumulative hazard function* at time t then is

$$\Lambda(t) = \int_0^t \lambda(u) du = \int_0^t -d \log S(u) = -\log S(t), \quad (9.2)$$

from which it follows that

$$S(t) = \exp[-\Lambda(t)]. \quad (9.3)$$

Numerous parametric survival distributions and corresponding hazard functions and event probability distributions can be used to describe the experience in a population. The simplest instance is the *exponential* survival distribution, in which it is assumed that the hazard function is constant over time, or $\lambda(t) = \lambda \forall t$, in which case $\Lambda(t) = \lambda t$, $S(t) = e^{-\lambda t}$, $F(t) = 1 - e^{-\lambda t}$, and $f(t) = \lambda e^{-\lambda t}$. Other parametric functions are explored as problems.

In general, no one parametric distribution will apply to every population. Thus, common methods of survival analysis are nonparametric or distribution free. These include the Kaplan-Meier estimate of the underlying survival distribution, the family of generalized Mantel-Haenszel tests for the comparison of the survival distributions of two or more groups, and the proportional hazards model for the assessment of the effects of covariates on the risk or hazard function of events over time.

9.1.2 Censoring at Random

If all subjects were followed until the time of the event such as mortality or failure, then it would be a simple matter to estimate the parameters of a parametric model from the corresponding likelihood. With right-censored observations, however, the full likelihood is a function of both the failure time distribution and the distribution of the right-censored times. Assume that the event-time T distribution has parameter θ , where the probability density function, hazard function, and survival distribution are designated as $f(t; \theta)$, $\lambda(t; \theta)$, and $S(t; \theta)$, respectively. Also, assume that the censoring times C have a cumulative distribution function $G(c; \phi)$ and probability density function $g(c; \phi)$. The assumption of random censoring is equivalent to the assumption that the two sets of parameters θ and ϕ are distinct.

If the i th subject is censored ($\delta_i = 0$) we observe $t_i = c_i$. Then for any event-time and censoring distributions, the likelihood of a sample of N observations $\{t_i, \delta_i\}$ under random censoring is

$$L(\theta, \phi) = \prod_{i=1}^N \{f(t_i; \theta) [1 - G(t_i; \phi)]\}^{\delta_i} \{g(t_i; \phi) S(t_i; \theta)\}^{1-\delta_i}, \quad (9.4)$$

where an individual with the event at time t_i is also not censored by that time, and an individual censored at time t_i has also survived to that time. Thus, the likelihood can be factored as

$$\begin{aligned} L(\theta, \phi) &= \prod_{i=1}^N f(t_i; \theta)^{\delta_i} S(t_i; \theta)^{1-\delta_i} \prod_{i=1}^N [1 - G(t_i; \phi)]^{\delta_i} g(t_i; \phi)^{1-\delta_i} \\ &= L(\theta)L(\phi) \propto L(\theta). \end{aligned} \quad (9.5)$$

Thus, the full likelihood is proportional to the likelihood associated with the event-time distribution $L(\theta)$, which implies that inferences regarding the event-time distribution may be obtained without the need to simultaneously specify or consider the censoring distribution.

When the data consist of observations from two or more groups or strata of subjects, the assumption of censoring at random then specifies that the observations within each group or strata are censored at random, but the random mechanisms and the extent of censoring may differ between groups. The resulting likelihood is then the product of the group- or stratum-specific event-time likelihoods.

9.1.3 Kaplan-Meier Estimator

Kaplan and Meier (1958) described a *product limit estimator* of the underlying survival distribution without assuming any particular parametric form. Under random censoring, the likelihood of the sample is

$$L \propto \prod_{i=1}^N f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^N \lambda(t_i)^{\delta_i} S(t_i). \quad (9.6)$$

Thus, individuals who experience the event contribute the term $f(t)$ to the likelihood at the event time, whereas censored observations contribute the term $S(t)$ at the time of censoring.

Consider a sample of N observations (t_i, δ_i) , $i = 1, \dots, N$, in which events are observed at J distinct event times $t_{(1)} < t_{(2)} < \dots < t_{(J)}$ and that the underlying survival distribution is a step function with points of discontinuity at these event times. The left-hand limit $F(t^-) = \lim_{\epsilon \downarrow 0} F(t - \epsilon) = \lim_{u \uparrow t} F(u)$ designates the value of the *cdf* immediately prior to a "jump" or discontinuity at time t . Then the probability of an event at time $t_{(j)}$ is $f(t_{(j)}) = F(t_{(j)}) - F(t_{(j)}^-)$. Thus, $f(t_{(j)})$ also equals the size of the drop in the survival function $f(t_{(j)}) = S(t_{(j)}^-) - S(t_{(j)})$, where $S(t_{(j)}^-) = P(T \geq t_{(j)})$ is the probability of being at risk at $t_{(j)}$ or survival up to $t_{(j)}$ and $S(t_{(j)}) = P(T > t_{(j)})$ is the probability of surviving beyond t_j . Since $f(t_{(j)})$ is the probability of survival to t_j and an event at t_j , then

$$f(t_{(j)}) = P[\text{event at } t_{(j)} \mid T \geq t_{(j)}] P[T \geq t_{(j)}], \quad (9.7)$$

which can be expressed as $f(t_{(j)}) = \pi_j S(t_{(j)}^-)$, where the conditional probability of an event at $t_{(j)}$ is

$$\pi_j = \lim_{\Delta t \downarrow 0} P[t_{(j)}^- < T \leq t_{(j)}^- + \Delta t \mid T > t_{(j)}^-], \quad (9.8)$$

and its complement is the continuation probability

$$1 - \pi_j = P[T > t_{(j)} \mid T > t_{(j)}^-] = P[T > t_{(j)} \mid T \geq t_{(j)}]. \quad (9.9)$$

Since $S(t)$ is a decreasing step function, then $f(t_{(j)}) = \pi_j S(t_{(j)}^-) = \pi_j S(t_{(j-1)})$.

By construction the observed event times $\{t_i, \delta_i = 1\} \in \{t_{(1)}, \dots, t_{(J)}\}$. Let $n_j = \#\{t_i \geq t_{(j)}\}$ denote the number of subjects at risk, or still under observation, at time $t_{(j)}$. Then let $d_j = \#\{t_i = t_{(j)}, \delta_i = 1\}$ denote the number of events observed at time $t_{(j)}$ among the n_j subjects at risk at time $t_{(j)}$. Also, let w_j denote the number of observations that are right censored at times after the j th event time but prior to the $(j+1)$ th time, or $w_j = \#\{t_i, \delta_i = 0\} \in [t_{(j)}, t_{(j+1)})$, so that $n_{j+1} = n_j - d_j - w_j$. Thus, observations that are right censored at event time $t_{(j)}$ are considered to be at risk at that time and are removed from the risk set immediately thereafter. Then the likelihood can be expressed as

$$L(\pi_1, \dots, \pi_N) \propto \prod_{j=1}^J \pi_j^{d_j} S(t_{(j-1)})^{d_j} S(t_{(j)})^{w_j}, \quad (9.10)$$

where $t_{(0)} = 0$ and $S(t_{(0)}) = 1$.

An essential property of an event-time process is that in order to survive event-free beyond time $t_{(j)}$ requires that the subject survive event-free beyond $t_{(j-1)}$, and $t_{(j-2)}$, and so on. Thus,

$$\begin{aligned} S(t_{(j)}) &= P(T > t_{(j)}) = P(T > t_{(j)} \mid T > t_{(j-1)})P(T > t_{(j-1)}) \\ &= P(T > t_{(j)} \mid T > t_{(j-1)})P(T > t_{(j-1)} \mid T > t_{(j-2)})P(T > t_{(j-2)}), \end{aligned} \quad (9.11)$$

and so on. Note, however, that

$$P(T > t_{(j)} \mid T > t_{(j-1)}) = P(T > t_{(j)} \mid T \geq t_{(j)}) = (1 - \pi_j) \quad (9.12)$$

so that

$$\begin{aligned} S(t_{(j)}) &= (1 - \pi_j) S(t_{(j-1)}) \\ &= (1 - \pi_j)(1 - \pi_{j-1}) \cdots (1 - \pi_2)(1 - \pi_1). \end{aligned} \quad (9.13)$$

After expanding the likelihood in this manner and simplifying (see Problem 9.8), it follows that

$$L(\pi_1, \dots, \pi_J) \propto \prod_{j=1}^J \pi_j^{d_j} (1 - \pi_j)^{n_j - d_j}, \quad (9.14)$$

with log likelihood

$$\ell(\pi_1, \dots, \pi_J) = \sum_{j=1}^J \{d_j \log(\pi_j) + (n_j - d_j) \log(1 - \pi_j)\}. \quad (9.15)$$

It then follows that the maximum likelihood estimates of the event probabilities are provided by the solution to the estimating equations

$$\frac{\partial \ell(\pi_1, \dots, \pi_J)}{\partial \pi_j} = \frac{d_j}{\pi_j} - \frac{n_j - d_j}{1 - \pi_j} = 0; \quad 1 \leq j \leq J, \quad (9.16)$$

from which the *MLE* of the j th conditional probability is

$$\begin{aligned} \hat{\pi}_j &= \frac{d_j}{n_j} = p_j \\ 1 - \hat{\pi}_j &= \frac{n_j - d_j}{n_j} = q_j. \end{aligned} \quad (9.17)$$

Thus, Kaplan and Meier (1958) showed that the generalized maximum likelihood estimate of the underlying survival function is provided by

$$\hat{S}(t) = \prod_{j=1}^J \left(\frac{n_j - d_j}{n_j} \right)^{I[t_{(j)} \leq t]} = \prod_{j=1}^J q_j^{I\{t_{(j)} \leq t\}} = \prod_{j:t_{(j)} \leq t} q_j \quad (9.18)$$

for values $0 \leq t \leq t_{(J)}$ and is undefined for values $t > t_{(J)}$, $I\{\cdot\}$ being the indicator function.

Under random censoring, it can then be shown using classical methods (Peterson, 1977), or using counting processes (Aalen, 1978, Gill, 1980), that $\lim_{n \rightarrow \infty} \widehat{S}(t) \xrightarrow{P} S(t)$ whatever the form of the underlying distribution. Clearly, $\widehat{S}(t)$ is a step function that describes the probability of surviving beyond time t . Its complement is the *cumulative incidence function*, which describes the probability of the event occurring up to time t . As the number of observations increases, and thus the number of events also increases, the number of steps increases and the intervals between steps become infinitesimally small, thus converging to the true survival distribution, whatever its form.

The estimator is obtained as a product of successive survival proportions, each being the conditional probability of surviving beyond an instant of time given survival up to that point in time. Thus, the assumption of censoring at random implies that the survival experience of those at risk at any event time $t_{(j)}$ applies to all observations in the population, including those censored prior to $t_{(j)}$; or that the subset of n_j observations at risk at $t_{(j)}$ is a random subset of the initial cohort who actually survive to $t_{(j)}$. Alternatively, random censoring implies that the estimate of the survival experience beyond any point in time is not biased by prior censored observations, or that the censored observations are subject to the same hazard function as those who continue follow-up.

These successive conditional event proportions $\{p_j\}$ clearly are not statistically independent. However, it is readily shown that these successive proportions are indeed uncorrelated; see Problem 9.8.4. Because the vector of proportions is asymptotically distributed as multivariate normal, this implies that these proportions are asymptotically conditionally independent given the past sequence of events and censored observations, that is, conditionally on the numbers at risk at the times of the preceding events.

Because the survival function estimate is a product of probabilities, it is more convenient to use the log survival function to obtain expressions for the variance of the estimate, where

$$\log[\widehat{S}(t)] = \sum_{j=1}^J I[t_{(j)} \leq t] \log[q_j] = \sum_{j:t_{(j)} \leq t} \log[q_j]. \quad (9.19)$$

Using the δ -method, the variance is

$$V(\log[\widehat{S}(t)]) = \sum_{j:t_{(j)} \leq t} \frac{\pi_j}{n_j(1 - \pi_j)}, \quad (9.20)$$

as is readily demonstrated in a Problem. Substituting $\widehat{\pi}_j$ from (9.17) yields the estimated variance

$$\widehat{V}(\log[\widehat{S}(t)]) = \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}. \quad (9.21)$$

Again, applying the δ -method yields Greenwood's (1926) variance estimate of the survival probability at any time $t \leq t_{(J)}$, which is expressed as

$$\widehat{V}[\widehat{S}(t)] = \widehat{S}(t)^2 \left[\sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)} \right]. \quad (9.22)$$

Greenwood's estimator can be used to provide symmetric confidence bands for the survival function. However, these confidence limits are not bounded by (0,1). Asymmetric confidence limits so bounded are obtained using the complementary log-log transformation, with corresponding variance again obtained by the δ -method as

$$\widehat{V} \left[\log \left(-\log[\widehat{S}(t)] \right) \right] = \frac{1}{\left(\log[\widehat{S}(t)] \right)^2} \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}. \quad (9.23)$$

This yields asymmetric $(1 - \alpha)$ -level confidence limits for $S(t)$ of the form

$$\widehat{S}(t)^{\exp[\pm Z_{1-\alpha/2} \sqrt{\widehat{V}[\log(-\log[\widehat{S}(t)])]}]}, \quad (9.24)$$

(cf. Collett, 1994). Alternatively, asymmetric confidence limits could be obtained using the logit of the survival probability (see Problem 9.8.9).

Finally, note that since $S(t) = \exp[-\Lambda(t)]$, then the estimate of the survival function at an event time also provides an estimate of the cumulative hazard as simply $\widehat{\Lambda}(t_{(j)}) = -\log[\widehat{S}(t_{(j)})]$. From (9.21) the estimated variance is simply

$$\widehat{V} \left[\widehat{\Lambda}(t_{(j)}) \right] = \sum_{\ell=1}^j \frac{d_\ell}{n_\ell(n_\ell - d_\ell)}. \quad (9.25)$$

In what follows we designate the Kaplan-Meier estimate of the survival and cumulative hazard function as $\widehat{S}_{KM}(t)$ and $\widehat{\Lambda}_{KM}(t)$, to distinguish them from the Nelson-Aalen estimates, which are based on estimation of the cumulative hazard function directly.

9.1.4 Estimation of the Hazard Function

When event times are observed continuously over time, one approach to estimation of the hazard function is to adopt a piecewise exponential model where we assume that the hazard function is piecewise constant between the successive event times, or $\lambda(t) = \lambda_{(j)}$ for $t \in (t_{(j-1)}, t_j]$ for the j th interval. The probability of the event over a small interval of time is $S(t_{(j-1)}) - S(t_j) = \exp[-\lambda_{(j)}t_{(j-1)}] - \exp[-\lambda_{(j)}(t_j)] \doteq \lambda_{(j)}(t_j - t_{(j-1)})$ since $e^{-\varepsilon} \doteq 1 - \varepsilon$ for small $\varepsilon \downarrow 0$. This probability is intuitively estimated by p_j in (9.17), which is simply the number of events divided by the number at risk at time $t_{(j)}$. Thus, the hazard function over the interval prior to the j th event time is estimated as

$$\widehat{\lambda}_{(j)} = \frac{p_j}{t_{(j)} - t_{(j-1)}}. \quad (9.26)$$

Then the cumulative hazard up to that time $t_{(j)}$ is estimated as

$$\widehat{\Lambda}_{NA}(t_{(j)}) = \sum_{\ell=1}^j \widehat{\lambda}_{(\ell)}(t_{(\ell)} - t_{(\ell-1)}) = \sum_{\ell=1}^j p_\ell. \quad (9.27)$$

As described in Section 9.7, (9.27) is the *Nelson-Aalen estimator* $\widehat{\Lambda}_{NA}(t_{(j)})$ after Nelson (1969, 1972) and Aalen (1978), who provided estimates of the intensity of a

counting process. The Nelson-Aalen estimate of the cumulative hazard function then yields the Nelson-Aalen estimate of the survival function at time $t_{(j)}$ as

$$\widehat{S}_{NA}(t_{(j)}) = \exp \left[-\widehat{\Lambda}_{NA}(t_{(j)}) \right], \quad (9.28)$$

which was also proposed by Altschuler (1970).

Different expressions apply to discrete-time observations, such as for the actuarial lifetable described in Section 9.2.1. These are presented in many standard texts.

Since the successive proportions are uncorrelated, then it readily follows that the estimated variances are

$$\widehat{V}(\widehat{\lambda}_{(j)}) = \frac{p_j(1-p_j)}{n_j(t_{(j)} - t_{(j-1)})^2}, \quad (9.29)$$

$$\widehat{V} \left[\widehat{\Lambda}_{NA}(t_{(j)}) \right] = \sum_{\ell=1}^j \frac{p_\ell(1-p_\ell)}{n_\ell}, \quad (9.30)$$

and

$$\widehat{V} \left[\widehat{S}_{NA}(t_{(j)}) \right] = \widehat{S}_{NA}(t_{(j)})^2 \sum_{\ell=1}^j \frac{p_\ell(1-p_\ell)}{n_\ell}. \quad (9.31)$$

From 9.28, asymmetric confidence limits on the survival probabilities can be obtained from those on the estimated cumulative hazard $\widehat{\Lambda}_t$. Also, since

$$\widehat{V} \left[\log \left(-\log[\widehat{S}_{NA}(t)] \right) \right] = \widehat{V} \left[\widehat{\Lambda}_{NA}(t) \right] \left(\log[\widehat{S}_{NA}(t)] \right)^{-2}, \quad (9.32)$$

then asymmetric confidence limits for $S(t)$ may be obtained from (9.24) upon substituting the corresponding Nelson-Aalen estimates.

In Problem 9.8.11 it is shown that as $N \rightarrow \infty$, the Nelson-Aalen and Kaplan-Meier estimators of the survival function are asymptotically equivalent, and that the Nelson-Aalen and Kaplan-Meier estimators of the variance of the cumulative hazard are also asymptotically equivalent. In Section 9.7.2 we show that the Nelson-Aalen estimators can be applied to the intensity of a counting process for recurrent events. Smoothed estimates of the intensity described therein may also be applied to the Nelson-Aalen estimate of the hazard function for survival data.

9.1.5 Comparison of Survival Probabilities for Two Groups

In many instances we wish to compare the event-time distributions between two independent groups of subjects. We now consider the comparison of two groups at a specific point in time using the Kaplan-Meier estimate. In Section 9.3 we describe tests for the difference between hazard functions and survival curves over time.

The Greenwood variance estimator provides a large sample confidence interval for the difference between the survival probabilities of the two groups at any point in time ($t > 0$), say $\widehat{S}_1(t) - \widehat{S}_2(t)$, with $\widehat{V}[\widehat{S}_1(t) - \widehat{S}_2(t)] = \widehat{V}[\widehat{S}_1(t)] + \widehat{V}[\widehat{S}_2(t)]$, where $\widehat{S}_i(t)$ and $\widehat{V}[\widehat{S}_i(t)]$ are obtained from the observations in each group separately ($i = 1, 2$). Alternatively, asymmetric confidence limits on the ratio of the survival probabilities

at time t can be obtained using the log transformation since $\log \widehat{S}_1(t) - \log \widehat{S}_2(t) = \log[\widehat{S}_1(t)/\widehat{S}_2(t)]$ with estimated variance $\widehat{V}[\log \widehat{S}_1(t)] + \widehat{V}[\log \widehat{S}_2(t)]$ from (9.21).

Likewise, asymmetric confidence limits on the ratio of the cumulative hazards can be obtained using the complementary log-log transformation since

$$\log\{-\log \widehat{S}_1(t)\} - \log\{-\log \widehat{S}_2(t)\} = \log[\widehat{\Lambda}_1(t)/\widehat{\Lambda}_2(t)] \quad (9.33)$$

with estimated variance $\widehat{V}[\log\{-\log \widehat{S}_1(t)\}] + \widehat{V}[\log\{-\log \widehat{S}_2(t)\}]$ from (9.23). Asymmetric confidence limits may also be obtained for the survival odds ratio at time t using the difference between the logits of the survival probabilities.

These developments also provide a large sample test of the difference between the survival probabilities of two groups at a specific point in time, say t^* . Under the null hypothesis $H_0: S_1(t) = S_2(t)$, the estimated survival function from the combined groups, say $\widehat{S}_*(t)$, provides a consistent estimator of the common survival function. This pooled estimate is computed from (9.18) with estimated conditional event probabilities $\widehat{\pi}_{\bullet j} = (d_{1j} + d_{2j})/(n_{1j} + n_{2j})$ for $1 \leq j \leq J$. Then for the i th group ($i = 1, 2$), from (9.22) the estimated variance of $S_i(t^*)$ at the specified time t^* under the null hypothesis is expressed as

$$\widehat{V}_0[\widehat{S}_i(t^*)] = \widehat{S}_*(t^*)^2 \sum_{j: t_{(j)} \leq t^*} \frac{\widehat{\pi}_{\bullet j}}{n_{ij}(1 - \widehat{\pi}_{\bullet j})}, \quad (9.34)$$

where n_{ij} is the number at risk in the i th group at the j th event time. Thus, a large sample test of equality of the survival probabilities at time t^* is provided by

$$X^2 = \frac{[\widehat{S}_1(t^*) - \widehat{S}_2(t^*)]^2}{\widehat{V}_0[\widehat{S}_1(t^*)] + \widehat{V}_0[\widehat{S}_2(t^*)]}, \quad (9.35)$$

where asymptotically X^2 is distributed as chi-square on 1 df .

An alternative naive approach is to simply compute the test for two proportions, ignoring the presence of censored observations and variable durations of follow-up. We then wish to test $H_0: \pi_1 = \pi_2 = \pi$, where π_i denotes the aggregate probability of an event among the n_i observations in the i th group. The simple test for two proportions in Section 2.6.2 assumes that $E(\delta_{ij}) = \pi$ for all subjects ($j = 1, \dots, n_i$; $i = 1, 2$), where δ_{ij} is the Bernoulli variable for the occurrence of the event versus being censored. This would be appropriate if all subjects were to be followed for a common fixed duration, or were right censored at a common fixed time if the event had not yet occurred. However, this is no longer appropriate when the set of exposure or right censoring times varies among subjects. For the ij th subject, let e_{ij} denote the possible exposure time, meaning the time at which follow-up will be terminated if the event has not yet occurred. Then each subject has a different implied probability of the event $E(\delta_{ij}|e_{ij}) = \pi_{ij}$ given the exposure time e_{ij} . Clearly, if the survival function $S_i(t)$ is declining sharply over the range of the e_{ij} , then the variance of the simple proportions test will be affected.

Table 9.1 Numbers at risk (n_j) in both groups individually and combined at each distinct event time $t_{(j)}$, numbers of events (d_j) at that time, and number lost to follow-up or right censored (w_j) during the interval $(t_{(j)}, t_{(j+1)})$.

$t_{(j)}$	n_j	d_j	w_j	n_{1j}	d_{1j}	w_{1j}	n_{2j}	d_{2j}	w_{2j}
1	85	1	5	23	0	2	62	1	3
2	79	2	3	21	0	0	58	2	3
3	74	2	1	21	0	0	53	2	1
4	71	1	5	21	0	4	50	1	1
5	65	5	1	17	1	1	48	4	0
6	59	1	1	15	0	0	44	1	1
7	57	1	1	15	0	0	42	1	1
8	55	6	0	15	1	0	40	5	0
9	49	2	0	14	0	0	35	2	0
10	47	1	1	14	0	0	33	1	1
11	45	7	1	14	2	0	31	5	1
12	37	3	2	12	2	0	25	1	2
13	32	2	2	10	1	1	22	1	1
14	28	0	2	8	0	2	20	0	0
15	26	0	1	6	0	0	20	0	1
16	25	1	0	6	0	0	19	1	0
17	24	1	1	6	0	0	18	1	1
18	22	0	2	6	0	0	16	0	2
19	20	2	0	6	1	0	14	1	0
20	18	1	0	5	0	0	13	1	0
21	17	4	1	5	1	0	12	3	1
22	12	2	1	4	0	0	8	2	1
26	9	1	0	4	0	0	5	1	0
28	8	0	1	4	0	0	4	0	1
29	7	1	0	4	0	0	3	1	0
30	6	1	1	4	0	1	2	1	0
34	4	1	0	3	1	0	1	0	0
55	3	1	0	2	0	0	1	1	0
84	2	1	0	2	1	0	0	0	0
88	1	1	0	1	1	0	0	0	0

Since the sample proportion in the i th group is $p_i = \sum_{j=1}^{n_i} \delta_{ij} / n_i$, then,

$$E(p_i) = \frac{\sum_j E(\delta_{ij} | e_{ij})}{n_i}, \quad (9.36)$$

where $E(\delta_{ij} | e_{ij}) = F_i(e_{ij})$. Thus,

$$V_c(p_i) = \frac{\sum_j V(\delta_{ij} | t_{ij})}{n_i^2} = \frac{\sum_{ij} F_i(t_{ij}) S_i(t_{ij})}{n_i^2}. \quad (9.37)$$

In Problem 9.9 it is then shown that

$$V_c(p_i) \leq V(p_i) = \pi_i(1 - \pi_i)/n_i, \quad (9.38)$$

with equality when $F(e_{ij})$ is nearly constant for all e_{ij} (Lachin et al., 1992). Therefore, the ordinary proportions test in (2.83) is flawed in an extended follow-up study with differential exposure among the subjects because the ordinary binomial variance is incorrect. This test, therefore, should not be used.

Example 9.1 *Squamous Cell Carcinoma*

Lagakos (1978) presents data on the time to spread of disease among patients with squamous cell carcinoma who received either of two treatments: A (experimental, group 1) versus B (control, group 2). Here we consider only the subset of patients who were nonambulatory on entry into the study. The complete data set is described later in Example 9.6 and can be obtained online (see the Preface, reproduced with permission). Patients were assigned treatment B versus A in a 3:1 allocation. In treatment group A , 12 of 23 patients (52%) failed versus 40 of the 62 patients (65%) in treatment group B .

Table 9.1 presents the times at which spread of the disease occurred along with the numbers still at risk at each event time and the numbers of events at that time. From this, the conditional probabilities of survival beyond each point in time are computed (q_j), as is the estimate of the survival probabilities.

From these entries the estimated conditional event probabilities (p_j) and continuation probabilities (q_j) are obtained, from which the estimated survival curves are obtained for both groups combined, along with the Greenwood estimated standard error and the piecewise constant hazard function and its standard error. These are presented in Table 9.2 for the combined group.

Figure 9.1 then presents a plot of the estimated event-free survival function and the asymmetric 95% confidence limits obtained using the standard error of the $\log(-\log)$ survival function. Table 9.3 presents the estimated survival and hazard functions for each group separately.

Figure 9.2 then presents a plot of the estimated event-free survival function in each group. There is a small increase in the probability of remaining free of the spread of cancer, or a small reduction in risk, with the experimental drug treatment. The apparent widening of the difference in survival functions beyond 20 weeks is highly unreliable because of the small numbers at risk in each group, especially the control group, as indicated by the large standard errors for $\hat{S}(t)$ in Table 9.3.

Suppose that the objective of the study was to compare the "survival" probabilities of remaining free of spread of disease beyond 24 weeks so that the event-free probabilities at this fixed point in time are the only quantities of interest. The estimated probabilities in each group are $\hat{S}_1(24) = 0.376$ and $\hat{S}_2(24) = 0.208$, with a difference of 0.169 with standard error 0.153. Using (9.34) the estimated variances under the null hypothesis are $\hat{V}_0[\hat{S}_1(24)] = 0.0129$ and $\hat{V}_0[\hat{S}_2(24)] = 0.00556$, which yields a pointwise chi-square test of the difference in remaining free of the spread of the disease for 24 weeks with value $X^2 = 1.540$ and $p \leq 0.22$.

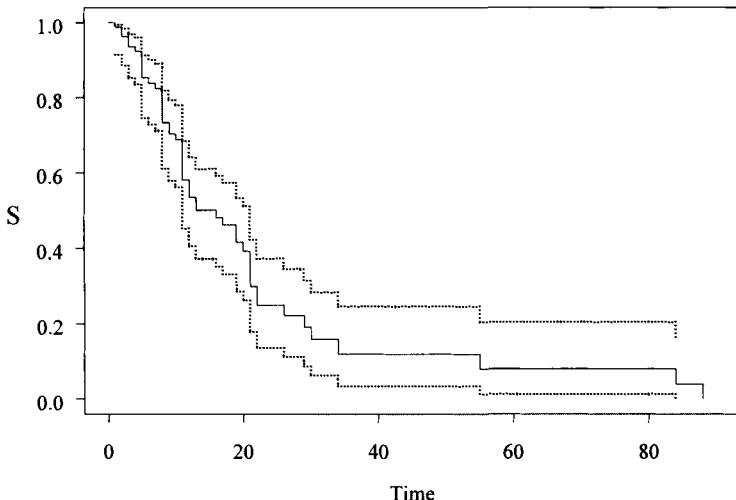
Table 9.2 Within the combined sample, time of each event $t_{(j)}$, number at risk (n_j), and number of events at that time (d_j), estimated continuation probability of survival beyond that time (q_j), Kaplan-Meier estimated probability $\hat{S}(t_{(j)})$, and Greenwood estimate of the $S.E.$, Nelson-Aalen estimate $\hat{\lambda}_j$ of the hazard over the interval $(t_{(j)}, t_{(j-1}]\text{, and its } S.E.$

$t_{(j)}$	$n_{(j)}$	d_j	q_j	$\hat{S}(t_{(j)})$	$S.E.[\hat{S}(t_{(j)})]$	$\hat{\lambda}_j$	$S.E.[\hat{\lambda}_j]$
1	85	1	0.988	0.988	0.012	0.012	0.012
2	79	2	0.975	0.963	0.021	0.025	0.018
3	74	2	0.973	0.937	0.027	0.027	0.019
4	71	1	0.986	0.924	0.030	0.014	0.014
5	65	5	0.923	0.852	0.041	0.077	0.033
6	59	1	0.983	0.838	0.043	0.017	0.017
7	57	1	0.982	0.824	0.045	0.018	0.017
8	55	6	0.891	0.734	0.053	0.109	0.042
9	49	2	0.959	0.704	0.055	0.041	0.028
10	47	1	0.979	0.689	0.056	0.021	0.021
11	45	7	0.844	0.582	0.060	0.156	0.054
12	37	3	0.919	0.535	0.061	0.081	0.045
13	32	2	0.937	0.501	0.061	0.062	0.043
16	25	1	0.960	0.481	0.062	0.013	0.013
17	24	1	0.958	0.461	0.063	0.042	0.041
19	20	2	0.900	0.415	0.064	0.050	0.034
20	18	1	0.944	0.392	0.065	0.056	0.054
21	17	4	0.765	0.300	0.064	0.235	0.103
22	12	2	0.833	0.250	0.062	0.167	0.108
26	9	1	0.889	0.222	0.061	0.028	0.026
29	7	1	0.857	0.190	0.060	0.048	0.044
30	6	1	0.833	0.159	0.058	0.167	0.152
34	4	1	0.750	0.119	0.055	0.062	0.054
55	3	1	0.667	0.079	0.049	0.016	0.013
84	2	1	0.500	0.040	0.037	0.017	0.012
88	1	1	0.000	0.000	.	0.250	0.000

In addition to the difference in event-free probabilities, other useful measures include the log ratio of event-free probabilities, or the difference in log event-free proportions; e.g., $\log[\hat{S}_1(24)/\hat{S}_2(24)] = 0.594$ with estimated standard error 0.490. Thus, the likelihood of survival beyond 24 weeks is $1.81 = \exp(0.594)$ times greater in group 1 than in group 2.

The $\log(-\log)$ of the event-free proportions, or the log cumulative hazards in each group are $\log[\hat{\lambda}_1(24)] = -0.0234$ and $\log[\hat{\lambda}_2(24)] = 0.452$, so that the difference, which equals the log cumulative hazard ratio, is -0.475 with estimated standard error 0.426. Under a proportional hazards model where the ratio of the hazards is assumed constant over time, the estimated hazard ratio over the 24 weeks of study is

Fig. 9.1 The Kaplan-Meier estimated probability of remaining free of spread of cancer and the 95% confidence limits at each point in time based on the log(-log) survival probabilities.



$\exp(-0.475) = 0.622$, indicating that the hazard of the event is approximately 38% lower in group 1 than in group 2.

Another useful summary is the survival odds ratio. The log odds of remaining event free in the experimental drug group is $\log[\hat{S}_1(24)/\hat{F}_1(24)] = -0.505$ and in the control group is $\log[\hat{S}_2(24)/\hat{F}_2(24)] = -1.338$. The log event-free odds ratio then is 0.834 with estimated standard error 0.715. This indicates that the odds of survival beyond 24 weeks is estimated to be $\exp(0.834) = 2.30$ times greater in group 1 than in group 2. The *S.E.* of this survival odds ratio could then be estimated using the results of Problem 9.8.9.

9.2 LIFETABLE CONSTRUCTION

Although the Kaplan-Meier estimate is derived assuming a discontinuous distribution, it provides an estimate of the survival probabilities for a continuous survival distribution at the times that events are observed to occur. This assumes that both the time of the event and the time of censoring are observed up to the instant of time for each subject. Usually, however, the closest one comes to continuous observation is when the events are recorded to the day in a long-term follow-up study, or to the hour in a short-term follow-up study. In this case, tied event times may be observed and some observations will have censored times that are tied with the event times. The

Table 9.3 Within each group, time of each event $t_{(j)}$, number at risk (n_{ij}), and number of events at that time (d_{ij}), estimated continuation probability of survival beyond that time (q_{ij}), Kaplan-Meier estimated probability $\hat{S}_i(t_{(j)})$, and Greenwood estimate of the $S.E.$, Nelson-Aalen estimate $\hat{\lambda}_{ij}$ of the hazard over the interval since the last event and its $S.E.$.

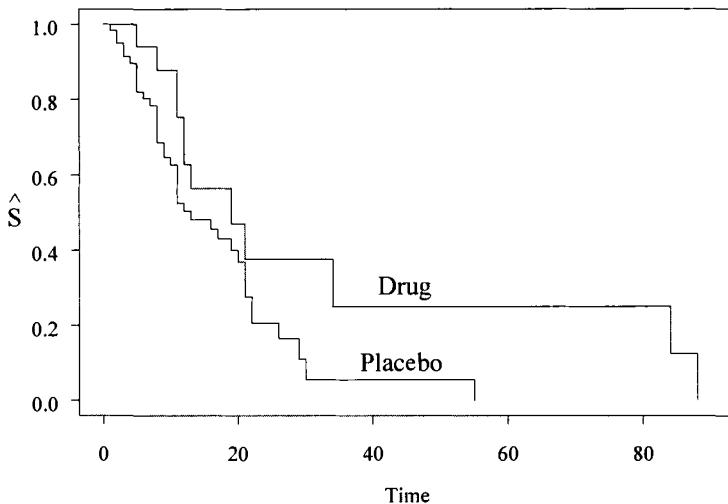
Experimentally Treated Group

$t_{(j)}$	n_{1j}	d_{1j}	q_{1j}	$\hat{S}_1(t_{(j)})$	$S.E.[\hat{S}_1(t_{(j)})]$	$\hat{\lambda}_{1j}$	$S.E.[\hat{\lambda}_{1j}]$
5	17	1	0.941	0.941	0.057	0.012	0.011
8	15	1	0.933	0.878	0.081	0.022	0.021
11	14	2	0.857	0.754	0.107	0.048	0.031
12	12	2	0.833	0.627	0.121	0.167	0.108
13	10	1	0.900	0.566	0.124	0.100	0.095
19	6	1	0.833	0.471	0.134	0.028	0.025
21	5	1	0.800	0.376	0.136	0.100	0.089
34	3	1	0.667	0.251	0.137	0.026	0.021
84	2	1	0.500	0.125	0.112	0.010	0.007
88	1	1	0.000	0.000	.	0.250	0.000

Control Treated Group

$t_{(j)}$	n_{2j}	d_{2j}	q_{2j}	$\hat{S}_2(t_{(j)})$	$S.E.[\hat{S}_2(t_{(j)})]$	$\hat{\lambda}_{2j}$	$S.E.[\hat{\lambda}_{2j}]$
1	62	1	0.984	0.984	0.016	0.016	0.016
2	58	2	0.966	0.950	0.028	0.034	0.024
3	53	2	0.962	0.914	0.037	0.038	0.026
4	50	1	0.980	0.896	0.040	0.020	0.020
5	48	4	0.917	0.821	0.051	0.083	0.040
6	44	1	0.977	0.802	0.054	0.023	0.022
7	42	1	0.976	0.783	0.056	0.024	0.024
8	40	5	0.875	0.685	0.064	0.125	0.052
9	35	2	0.943	0.646	0.066	0.057	0.039
10	33	1	0.970	0.627	0.067	0.030	0.030
11	31	5	0.839	0.526	0.070	0.161	0.066
12	25	1	0.960	0.505	0.070	0.040	0.039
13	22	1	0.955	0.482	0.070	0.045	0.044
16	19	1	0.947	0.456	0.071	0.018	0.017
17	18	1	0.944	0.431	0.071	0.056	0.054
19	14	1	0.929	0.400	0.073	0.036	0.034
20	13	1	0.923	0.369	0.073	0.077	0.074
21	12	3	0.750	0.277	0.072	0.250	0.125
22	8	2	0.750	0.208	0.069	0.250	0.153
26	5	1	0.800	0.166	0.066	0.050	0.045
29	3	1	0.667	0.111	0.063	0.111	0.091
30	2	1	0.500	0.055	0.050	0.500	0.354
55	1	1	0.000	0.000	.	0.040	0.000

Fig. 9.2 The Kaplan-Meier estimated probability of remaining free of spread of cancer in the experimentally treated and control groups.



usual convention is that any censored times that are also tied with an event time are included as being at risk at that time and are then censored or removed from the risk set immediately after the event time. This convention was employed in the analysis of Example 9.1. Modifications are necessary, however, when the event times are either grouped or are truly discrete.

9.2.1 Discrete Distributions: Actuarial Lifetable

In many studies the exact time of the events are unknown. Rather, the events are known to have occurred within some interval of time. For example, in a population lifetable, mortality is recorded up to the year of age, not the day. In this case the data are observed in *discrete or grouped time*. Likewise, in a follow-up study, events may be recorded up to the year of follow-up. In such cases, it is common to employ an *actuarial lifetable*, first described in the context of a follow-up study by Cutler and Ederer (1958). This is often termed simply the *lifetable method*. This method is appropriate when the same interval of time applies to all subjects. Herein, we consider factors that commonly arise in a discrete-time follow-up study.

In the more general case, observations may be known to have occurred only within intervals of time that vary from subject to subject. Such observations are *interval censored* (see Section 9.6.2).

In an actuarial lifetable, it is assumed that the events occur within fixed intervals of time and that censoring of follow-up also occurs within intervals of time. This can be viewed as converting continuous-time observations into *grouped time*. Here we assume that the event or censoring is observed to have occurred within one of a contiguous sequence of intervals $A_j = (\tau_{j-1}, \tau_j]$. Within the j th interval, let r_j be the number still at risk at the beginning of the interval, that is, event-free and under follow-up at τ_{j-1} , where $r_1 = N$ is the size of the cohort at the beginning of the study. Also, let d_j be the number of events that occur during the j th interval; and let w_j be the number of *study exits* or randomly censored (unbiased) observations during the interval. The ordinary actuarial estimate assumes that exits are known to be event-free (e.g., alive) at the time of exit. Usually, the exact time of exit is not known and thus we assume that the exits on average were followed for half the interval. Then the likelihood for the j th interval is

$$L(\pi_j) = \pi_j^{d_j} (1 - \pi_j)^{w_j/2} (1 - \pi_j)^{r_{j+1}}, \quad (9.39)$$

where $r_{j+1} = r_j - d_j - w_j$ (cf. Elandt-Johnson and Johnson, 1980). In Problem 9.8.3 it is shown that the resulting actuarial estimate of the conditional probability of experiencing an event during the interval is

$$p_j = \hat{\pi}_j = \frac{d_j}{n_j} \quad (9.40)$$

with denominator

$$n_j = r_j - w_j/2, \quad (9.41)$$

that is the estimated units or extent of exposure for the j th interval and is equivalent to the number at risk in the Kaplan-Meier estimate. The resulting actuarial estimator of the grouped survival function then is obtained as

$$\hat{S}[\tau_j] = \prod_{\ell=1}^j q_\ell = \prod_{\ell=1}^j \frac{n_\ell - d_\ell}{n_\ell}, \quad (9.42)$$

where the estimated continuation probability is $q_j = 1 - p_j$. The actuarial estimate $\hat{S}[\tau_j]$ thus provides an estimate of the probability of surviving beyond the j th interval, or $\hat{P}(T > \tau_j)$. The complement $1 - \hat{S}[\tau_j]$ provides an estimate of the cumulative probability of the event up through the j th interval, or $\hat{P}(T \leq \tau_j)$.

The estimate is of the same form as the Kaplan-Meier estimate, and thus the expressions for the variance of $\hat{S}[\tau_j]$, its log, and so forth, are similar to those presented in Section 9.1.3 upon substituting τ_j for t_j . Estimates of the hazard function within each interval and the cumulative hazard function for the actuarial lifetable are described in many general texts (cf. Marubini and Valsecchi, 1995).

9.2.2 Modified Kaplan-Meier Estimator

The actuarial estimate assumes that all subjects who exit during an interval are *known* to be event free at the time of exit. In many instances this may not apply, in which

consisting of the indices of all subjects in the original cohort still at risk at time $t_{(j)}$. Then the total probability of an event occurring at $t_{(j)}$ among all subjects in the risk set is

$$\sum_{\ell \in R(t_{(j)})} \lambda(t_{(j)} | \mathbf{x}_\ell) = \sum_{\ell \in R(t_{(j)})} \lambda_0(t_{(j)}) e^{\mathbf{x}'_\ell \boldsymbol{\beta}}. \quad (9.62)$$

Therefore, the conditional probability of an event occurring at $t_{(j)}$ in a subject with covariate vector $\mathbf{x}_{(j)}$ at that time is

$$\frac{\lambda(t_{(j)} | \mathbf{x}_{(j)})}{\sum_{\ell \in R(t_{(j)})} \lambda(t_{(j)} | \mathbf{x}_\ell)} = \frac{e^{\mathbf{x}'_{(j)} \boldsymbol{\beta}}}{\sum_{\ell \in R(t_{(j)})} e^{\mathbf{x}'_\ell \boldsymbol{\beta}}}. \quad (9.63)$$

This leads to the likelihood

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{j=1}^{d_*} \frac{e^{\mathbf{x}'_{(j)} \boldsymbol{\beta}}}{\sum_{\ell \in R(t_{(j)})} e^{\mathbf{x}'_\ell \boldsymbol{\beta}}}, \quad (9.64)$$

that is a function only of the times at which events occur. In fact, however, as is shown below, this is not a complete likelihood but rather is a partial likelihood, the complete likelihood involving other terms. Thus, the partial likelihood is designated as $\tilde{L}(\boldsymbol{\beta})$.

This likelihood can also be expressed in terms of individuals rather than event times. Using the joint observations (δ_i, t_i) , where δ_i is the indicator variable that denotes either the event or right censoring at time t_i , $i = 1, \dots, N$, Cox's partial likelihood is

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{i=1}^N \left[\frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\sum_{\ell \in R(t_i)} e^{\mathbf{x}'_\ell \boldsymbol{\beta}}} \right]^{\delta_i}, \quad (9.65)$$

where only those individuals with an event ($\delta_i = 1$) contribute to the likelihood. The partial likelihood can also be expressed as

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{i=1}^N \left[\frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\sum_{\ell=1}^N Y_\ell(t_i) e^{\mathbf{x}'_\ell \boldsymbol{\beta}}} \right]^{\delta_i} = \frac{\exp \left[\sum_{i=1}^N \delta_i \mathbf{x}'_i \boldsymbol{\beta} \right]}{\prod_{i=1}^N \left[\sum_{\ell=1}^N Y_\ell(t_i) e^{\mathbf{x}'_\ell \boldsymbol{\beta}} \right]^{\delta_i}}, \quad (9.66)$$

where $Y(t)$ refers to a time-dependent at-risk indicator variable to denote being at risk at time t . In the analysis of survival times, or the time of a single event,

$$Y_i(u) = \begin{cases} 1 : u \leq t_i \\ 0 : u > t_i \end{cases}, \quad u \geq 0. \quad (9.67)$$

Table 9.4 Number at risk (n) and number of events (d) at each annual evaluation in the intensive and conventional treatment groups of the secondary intervention cohort of the DCCT.

<i>Visit</i>	<i>Intensive</i>	<i>Conventional</i>		
t_j	n_{1j}	d_{1j}	n_{2j}	d_{2j}
1	325	18	316	19
2	307	15	297	22
3	290	9	275	17
4	281	7	258	16
5	274	7	242	10
6	204	7	167	5
7	129	2	96	3
8	73	2	52	4
9	57	2	33	3

9 years, averaging 6.5 years. Patients who were followed up to the termination of the study in the spring of 1993 who had not developed microalbuminuria were then *administratively censored* at that time.

Table 9.4 presents the numbers at risk in the i th group at the j th annual evaluation (n_{ij}) and the number of patients who had microalbuminuria for the first time at that visit (d_{ij}) among the 715 patients in the secondary intervention cohort of the study (see Section 1.5). The number of patients censored immediately after the j th evaluation in the i th group is $w'_{ij} = n_{i(j+1)} - n_{ij} + d_{ij}$. In this case, the number at risk is the number of annual renal evaluations conducted so that the numbers censored are transparent. Only two subjects were lost to follow-up (censored) during the first four years.

Table 9.5 then presents the resulting estimates of the survival function, the corresponding cumulative incidence function displayed in Figure 1.2, and the Greenwood estimated standard errors.

In both treatment groups there is a steady progression in the cumulative incidence of microalbuminuria. However, beyond two years of treatment, the risk in the intensive treatment group is substantially less than that in the conventional group, estimated to be 26% after nine years versus 42%.

9.2.3 SAS PROC LIFETEST: Survival Estimation

Some of the above analyses are provided by standard programs such as the SAS procedure LIFETEST. For continuous observations these programs compute the Kaplan-Meier estimate and its standard errors directly. For the squamous cell carcinoma data in Example 9.1, the SAS procedure LIFETEST with the following specification provides estimates of the survival function and its standard errors within each group as shown in Table 9.3:

Table 9.5 Estimated survival and cumulative incidence functions for developing microalbuminuria at each annual evaluation in the intensive and conventional treatment groups of the secondary intervention cohort of the DCCT.

<i>Visit</i>	<i>Intensive</i>			<i>Conventional</i>		
	<i>t_j</i>	\hat{S}_{1j}	\hat{F}_{1j}	<i>S.E.</i>	\hat{S}_{2j}	\hat{F}_{2j}
1	0.945	0.055	0.013	0.940	0.060	0.013
2	0.898	0.102	0.017	0.870	0.130	0.019
3	0.871	0.129	0.019	0.916	0.184	0.022
4	0.849	0.151	0.020	0.766	0.234	0.024
5	0.827	0.173	0.021	0.734	0.266	0.025
6	0.799	0.201	0.023	0.712	0.288	0.026
7	0.786	0.214	0.024	0.690	0.310	0.028
8	0.765	0.235	0.028	0.637	0.363	0.036
9	0.738	0.262	0.033	0.579	0.421	0.046

```
proc lifetest;
time time*delta(0); strata / group=treatment;
;
```

where the variable *time* is the survival/censoring time and censoring is indicated by the variable *delta=0*. The *strata* statement allows specification of strata (in this case none) and the group variable *treatment*. The program also provides plots of the survival and hazard functions, among other options.

However, with grouped or discrete data one must apply such programs with care because some programs, such as LIFETEST, employ different conventions from those employed herein. Throughout we take the view that the survival function is $S(t) = P(T > t)$, whereas LIFETEST defines the survival function as $S(t) = P(T \geq t)$. Thus, for grouped time data, the *j*th interval is defined herein as $A_j = (\tau_{j-1}, \tau_j]$, whereas LIFETEST defines this interval as $A_j = [\tau_{j-1}, \tau_j)$, closed on the left rather than the right. Thus, the user must be careful when using this or another program to construct the basic table of numbers at risk and numbers of events within an interval, such as for the DCCT nephropathy data in Example 9.2. In such cases, it is recommended that one should carefully construct this basic table before using a standard program. However, the output of the program will still appear somewhat different because of the manner in which the intervals are defined.

Example 9.3 Actuarial Lifetable in PROC LIFETEST

For example, the following is a hypothetical set of data from a study where survival was grouped into two-year intervals. The number entering the interval (r_j), the number of deaths during the interval (d_j), the number lost to follow-up and known to still be alive during the interval (w_j), the units of exposure (n_j) from (9.41),

the estimated probability of survival *beyond* each interval $\widehat{S}(\tau_j)$, and its estimated standard error are

$(\tau_{j-1}, \tau_j]$	r_j	d_j	w_j	n_j	$\widehat{S}(\tau_j)$	<i>S.E.</i>
(0, 2]	305	18	45	282.5	0.936	0.015
(2, 4]	242	12	22	231.0	0.888	0.019
(4, 6]	208	6	9	203.5	0.861	0.022

These data would be analyzed using the SAS procedure LIFETEST, using the following statements

```
Data one; input time freqn delta @@; cards;
2 18 1 2 45 0
4 12 1 4 22 0
6 6 1 6 9 0
8 0 1 8 193 0
;
PROC LIFETEST METHOD=act WIDTH=2;
    TIME time*delta(0); FREQ freqn;
```

Note the addition of an extra interval with the number censored after the study was concluded.

The SAS output would then include the following (extraneous material deleted):

Interval [Lower, Upper)	Number Failed	Number Censored	Effective Sample Size	Survival	Survival Standard Error
0	2	0	305.0	1.0000	0
2	4	18	282.5	1.0000	0
4	6	12	231.0	0.9363	0.0145
6	8	6	203.5	0.8876	0.0194
8	.	0	96.5	0.8615	0.0216

Because of the manner in which the intervals and the survival function are defined, the survival function estimates are offset by one interval compared to those computed directly from the data. Thus, the last extra interval is required in order to obtain the estimates of survival beyond six years using LIFETEST.

Thus, for other than continuous observations, the user should take care to properly construct the appropriate time variables for the censored and noncensored observations to ensure that the lifetable estimates are computed properly.

The computations herein are based on the author's program *survanal.sas* (available online, see the Preface) that provides macros that correspond to the conventions employed in the text.

9.3 FAMILY OF WEIGHTED MANTEL-HAENSZEL TESTS

9.3.1 Weighted Mantel-Haenszel Test

The Kaplan-Meier and actuarial estimates provide a description of the survival experience in a cohort over time. However, the simple test for the difference between groups in proportions surviving beyond a specific point in time in (9.35) leaves much to be desired because it does not provide a test of equality of the complete survival curves, or equivalently, the hazard functions, for two or more cohorts over time. Thus, another major advance was provided by Mantel (1966), when he reasoned that the Mantel-Haenszel test for a common odds ratio over independent strata also provides a test for differences between groups with respect to the pattern of event times, as represented by the underlying hazard function over time. Although the conditional proportions of events among those at risk within a cohort over time are not statistically independent, they are still uncorrelated and are asymptotically conditionally independent. Thus, Mantel suggested that the Mantel-Haenszel test statistic applied to survival times should converge in distribution to the central 1 *df* chi-square distribution. A general formal proof of Mantel's conjecture was later provided through application of the theory of martingales for counting processes by Aalen (1978) and Gill (1980); see also Harrington and Fleming (1982).

In Chapter 4 we showed that the Mantel-Haenszel test for multiple 2×2 tables may be differentially weighted over tables, the weights providing a test that is asymptotically fully efficient under a specific model, or efficient against a specific alternative hypothesis. So also, various authors have shown that a *weighted Mantel-Haenszel test* using a specific set of weights is asymptotically fully efficient against a specific alternative.

Let $t_{(j)}$ denote the j th ordered event time from the combined sample of two independent groups of sizes n_1 and n_2 initially at risk at time zero, $j = 1, \dots, J$, where J is the number of distinct event times, $J \leq d_*$, d_* being the total number of subjects with an event. Then, at the j th event time $t_{(j)}$ a 2×2 table can be constructed from the n_{ij} subjects still under follow-up and at risk of the event in the i th group ($i = 1, 2$). In the notation of stratified 2×2 tables introduced in Section 4.2.1, let a_j denote the number of events observed at $t_{(j)}$ in group 1 and let m_{1j} denote the total number of events in both groups at time $t_{(j)}$. Then the expected value of a_j under the null hypothesis H_0 : $S_1(t) = S_2(t) \forall t$, or equivalently, H_0 : $\lambda_1(t) = \lambda_2(t) \forall t$, is estimated to be $\widehat{E}(a_j|H_0) = n_{1j}m_{1j}/N_j$, where $N_j = n_{1j} + n_{2j}$ is the total number at risk at that time. The weighted Mantel-Haenszel test is then of the form

$$X_{WMH}^2 = \frac{\left(\sum_{j=1}^J w(t_{(j)}) [a_j - E(a_j|H_0)] \right)^2}{\sum_{j=1}^J w(t_{(j)})^2 V_c(a_j|H_0)}, \quad (9.43)$$

where $w(t_{(j)}) \neq 0$ for some j , and the conditional hypergeometric expectation $E(a_j|H_0)$ and variance $V_c(a_j|H_0)$ are as presented in Section 4.2.3:

$$E(a_j|H_0) = \frac{n_{1j}m_{1j}}{N_j}, \quad \text{and} \quad V_c(a_j|H_0) = \frac{n_{1j}n_{2j}m_{1j}m_{2j}}{N_j^2(N_j - 1)}. \quad (9.44)$$

Asymptotically, the statistic is distributed as χ^2 on 1 *df* under H_0 , and clearly, the size of the test asymptotically will be the desired level α regardless of the chosen weights.

This test is a weighted generalization of the Mantel-Haenszel test analogous to the Radhakrishna weighted Cochran-Mantel-Haenszel test presented in Section 4.7. In principle, the asymptotically most powerful test against a specific alternative could be derived by maximizing the asymptotic efficiency. In this setting, however, for continuous event times, as $N \rightarrow \infty$, then $J \rightarrow \infty$ and the sums becomes integrals. Schoenfeld (1981) presents the expression for the noncentrality parameter of the test with no ties and shows that, in general, the optimal weights for a weighted Mantel-Haenszel test are of the form

$$w(t) \propto \log[\lambda_1(t)/\lambda_2(t)], \quad (9.45)$$

proportional to the log hazard ratio over time under the alternative hypothesis. In cases where the alternative can be expressed in the form

$$\log[\lambda_1(t)/\lambda_2(t)] \cong g[F_0(t)], \quad (9.46)$$

or as some function of the cumulative event distribution function under the null hypothesis, Schoenfeld (1981) shows that the optimal weights are

$$w(t) = g[\widehat{F}(t)], \quad (9.47)$$

where $\widehat{F}(t) = 1 - \widehat{S}(t)$ and $\widehat{S}(t)$ is the Kaplan-Meier estimate of the survival distribution for the combined sample. These expressions allow the determination of the optimal weights against a specific alternative, or the alternative for which a specific set of weights is fully efficient.

9.3.2 Mantel-Logrank Test

The test with unit weights $w(t) = 1 \ \forall t$ yields the original unweighted Mantel-Haenszel test for multiple 2×2 tables applied to lifetables as suggested by Mantel (1966). When there are no tied event times ($J = d_*$), then $m_{1j} = 1$ for $j = 1, \dots, d_*$ and the test is equivalent to a rank-based test statistic using a permutational variance derived by Peto and Peto (1972), which they termed the *logrank test*. However, when there are tied event times, the Peto-Peto logrank test statistic differs from the Mantel-Haenszel test statistic, and its variance is greater. The Mantel-Haenszel test, therefore, is widely referred to as the logrank test, even though the test of Peto and Peto does not accommodate tied observations in the same manner. Cox (1972) also showed that the Mantel-Haenszel or logrank test (without ties) could also be derived as an efficient score test in a proportional hazards regression model with a single binary covariate to represent treatment group (see Problem 9.14). Thus, the test is also called the *Mantel-Cox test*.

Based on the work of Cox (1972), Peto and Peto (1972), and Prentice (1978), among many, the Mantel-logrank test with $w(t) = 1$ is asymptotically fully efficient

against a *proportional hazards* or *Lehmann alternative*, which specifies that

$$S_1(t) = S_2(t)^\phi \quad (9.48)$$

for some real constant ϕ . Thus,

$$\exp[-\Lambda_1(t)] = \exp[-\phi\Lambda_2(t)], \quad (9.49)$$

and it follows that

$$\begin{aligned} d\Lambda_1(t) &= \phi d\Lambda_2(t) \\ \lambda_1(t) &= \phi \lambda_2(t) \end{aligned} \quad (9.50)$$

or that the hazard functions are proportional over time. From (9.45), therefore, the test with constant or unit weights $w(t) = 1$ is fully efficient against this alternative.

Under this alternative, the constant ϕ is the hazard ratio over time, which is a general measure of the instantaneous relative risk between the two groups. As shown below, an estimate of the relative risk and its variance is conveniently provided by the Cox proportional hazards model. A location shift of an exponential or a Weibull distribution also satisfies this alternative.

9.3.3 Modified Wilcoxon Test

Forms of the weighted Mantel-Haenszel test have also been shown to be equivalent to a modified Wilcoxon rank test for censored event times. The *Peto-Peto-Prentice Wilcoxon test* uses weights equal to the Kaplan-Meier estimate of the survival function for the combined sample, $w(t) = \hat{S}(t)$, after asymptotically equivalent tests derived independently by Peto and Peto (1972) and Prentice (1978). This test provides greater weight to the early events, the weights diminishing as $\hat{S}(t)$ declines. A similar test, the *Gehan modified Wilcoxon test* (Gehan, 1965), was shown by Mantel (1967) to be a weighted Mantel-Haenszel test using the weights $w(t_{(j)}) = n_j$ at the j th event time. However, since these weights depend on both the event-time distribution and the censoring-time distribution (Prentice and Marek, 1979), it is not possible to describe a general condition under which this test is optimal.

Conversely, the Peto-Peto-Prentice Wilcoxon test is asymptotically fully efficient against a *proportional odds alternative* of the form

$$\frac{F_1(t)}{1 - F_1(t)} = \left[\frac{F_2(t)}{1 - F_2(t)} \right] \varphi, \quad (9.51)$$

which specifies that the odds of the event are proportional over time, or that the event odds ratio φ is constant over time. Note that this model also specifies that the survival odds ratio is $1/\varphi$. This is also called a *logistic alternative* because this relationship is satisfied under a location shift of a logistic probability distribution (see Problem 9.4.2). Under this model, Bennett (1983), among others, shows that the Prentice (1978) Wilcoxon test can be obtained as an efficient score test of the effect of a binary covariate for treatment group. Also, using Schoenfeld's result in (9.47), in Problem 9.10 it is shown that the optimal weights under the proportional odds model are $w(t) = \hat{S}(t)$.

9.3.4 G^ρ Family of Tests

To allow for a family of alternatives, Harrington and Fleming (1982) proposed that the Mantel-Haenszel test be weighted by a power of the estimated survival function $w(t) = \hat{S}(t)^\rho$ at time t . Because they used G^ρ to denote the resulting test statistic for given ρ , the family of tests has come to be known as the G^ρ family of tests. A related family of tests was proposed by Tarone and Ware (1977), using weights $w(t_{(j)}) = n_j^\rho$ based on the total number at risk at the j th event time. In each family, the value of ρ determines the weight assigned to events at different times and thus the alternative for which the test is asymptotically efficient. For $\rho = 0$ in the G^ρ or Tarone-Ware family, all events are weighted equally, yielding the original unweighted Mantel-logrank test. When $\rho = 1$, the resulting G^ρ test is the Peto-Peto-Prentice Wilcoxon test, whereas the Tarone-Ware test is Gehan's test. The Tarone-Ware weights, however, reflect both the survival and the censoring distributions. Thus, the Harrington-Fleming formulation is preferred.

The relative efficiency of various members of these families of tests have been explored, focusing principally on the relative efficiency of the logrank and Wilcoxon tests under specific alternatives. Tarone and Ware (1977) explored the efficiency of the Mantel-logrank test versus the Gehan-Wilcoxon test by simulation. Peto and Peto (1972) derived both the logrank and a modified Wilcoxon test based on the theory of optimal linear rank tests of Hájek and Šídák (1967). Prentice (1978) also showed that the logrank and a modified Wilcoxon test could be obtained as asymptotically efficient linear rank tests using scores derived from an accelerated failure-time model. Prentice and Marek (1979) and Mehrotra et al. (1982) showed that these linear rank tests are algebraically equivalent to a weighted Mantel-Haenszel test with censored observations. Harrington and Fleming (1982) also explored in detail the relative efficiency of members of the G^ρ family of tests based on the Aalen-Gill martingale representation of the underlying counting process, where these tests are also equivalent to the weighted Mantel-Haenszel tests shown in (9.43); see Section 9.7.3.

Clearly, if the distributions from the two groups satisfy proportional hazards then they cannot also satisfy proportional survival odds, and vice versa. Further, the efficiency of a test with specified weights, or a specified value of ρ , will depend on the extent to which the data support the particular alternative hypothesis for which that test is most efficient. Thus, if one applies a test with logrank (constant) weights and the hazard functions are not proportional, there will be a loss of power. Lagakos and Schoenfeld (1984) describe the loss in power for the Mantel or logrank test when the hazard ratios are not constant (nonproportional) and, in fact, may cross. Conversely, if Wilcoxon weights are used and the hazards are proportional, then likewise there will be a loss in power. These cases are analogous to the loss of efficiency incurred in the test for multiple independent 2×2 tables described in Section 4.8, when the test for one scale is chosen and the true alternative is that there is a constant difference over strata on a different scale.

As was the case for multiple 2×2 tables, it is inappropriate to conduct multiple tests and then cite that one with the smallest p -value. Rather, in order to preserve the size

of the test, the specific test, and the specific set of weights, must be specified a priori. Thus, one risks a loss of power or efficiency if the chosen test is suboptimal. As for the analysis of multiple 2×2 tables described in Section 4.9, one safeguard against this loss of power, while preserving the size of the test, is to employ the Gastwirth (1985) maximin-efficient robust test (*MERT*), which is obtained as a combination of the extreme pair in the family of tests. Apart from the fact that the weighted Mantel-Haenszel tests use the conditional hypergeometric variance, whereas the Radhakrishna tests use the unconditional product-binomial variance, the *MERT* in the two cases is identical; see Gastwirth (1985). Thus, the developments in Section 4.9.2 also apply to the computation and description of the *MERT* for the G^ρ family of tests.

Finally, one may also conduct a stratified-adjusted test of the difference between the lifetables of two (or more) groups. If the total sample of N subjects are divided among K independent strata ($k = 1, \dots, K$), then the stratified-adjusted weighted Mantel-Haenszel test is of the form

$$X_{\rho A}^2 = \frac{\left(\sum_{k=1}^K \sum_{j=1}^{J_k} w_k(t_{(j,k)}) [a_{(j,k)} - E(a_{(j,k)}|H_0)] \right)^2}{\sum_{k=1}^K \sum_{j=1}^{J_k} w_k(t_{(j,k)})^2 V_c(a_{(j,k)}|H_0)}, \quad (9.52)$$

where $t_{(j,k)}$ denotes the j th of the J_k event times observed among the n_k subjects within the k th stratum, $w_k(t_{(j,k)})$ is the weight within the k th stratum at time $t_{(j,k)}$, $a_{(j,k)}$ is the observed number of events in group 1 at that time, $E(a_{(j,k)}|H_0)$ is its expected value with respect to the observations from the k th stratum at risk at that time, and $V_c(a_{(j,k)}|H_0)$ is its conditional variance. For the G^ρ family, $w_k(t_{(j,k)}) = \hat{S}_k(t_{(j,k)})^\rho$ based on the estimated survival function within the k th stratum. Breslow (1975) presents a justification for the stratified logrank test ($\rho = 0$) under a proportional hazards model (see also Problem 9.14.5). However, the stratified Gehan and Peto-Peto-Prentice Wilcoxon tests with censored observations have not been formally studied.

9.3.5 Measures of Association

For the analysis of simple and stratified 2×2 tables, various measures of association were employed, including the odds ratio and relative risk. Mantel (1966) suggested that the Mantel-Haenszel stratified-adjusted estimate of the common odds ratio be used to summarize the differences between two survival curves, each event time representing a stratum. As shown above, however, the Mantel or logrank test is asymptotically efficient under a proportional hazards alternative. Thus, Anderson and Bernstein (1985) show that the Mantel-Haenszel stratified-adjusted estimate of the common relative risk expressed in (4.17) provides an estimate of the assumed common hazard ratio over time.

Peto et al. (1976) suggested that a score-based estimate of the hazard ratio (relative risk) be obtained as $\hat{\phi} = e^{\hat{\beta}_{RR}}$, where $\hat{\beta}_{RR} = T_L/\hat{V}(T_L)$ with estimated variance $\hat{V}(\beta_{RR}) = \hat{V}(T_L)^{-1}$, when the Mantel or logrank test in (9.43) with unit weights

$[w(t) = 1]$ is expressed as $X_L^2 = T_L^2/\hat{V}(T_L)$. This estimate is derived as a first-order approximation to the *MLE* under the Cox proportional hazards model described subsequently, in like manner to the derivation of the Peto estimate of the odds ratio for a 2×2 table described in Section 6.4 (see Problem 9.14.6). This provides an estimate of the nature of the treatment group difference for which the logrank test is optimal.

Similarly, the Peto-Peto-Prentice Wilcoxon test may be obtained as an efficient score test under a proportional odds model (Bennett, 1983). This suggests that an estimate of the log odds ratio of the cumulative event probability, or the negative log of the survival odds ratio, may be obtained as $\hat{\beta}_{OR} = T_{PW}/\hat{V}(T_{PW})$ with estimated variance $\hat{V}(\hat{\beta}_{OR}) = \hat{V}(T_{PW})^{-1}$ using the Peto-Peto-Prentice Wilcoxon statistic T_{PW} and its estimated variance $\hat{V}(T_{PW})$.

Example 9.4 *Squamous Cell Carcinoma (continued)*

For the data from Lagakos (1978) presented in Example 9.1, the commonly used weighted Mantel-Haenszel test statistics are

Test	Weights	Statistic	Variance	X^2	$p \leq$
Logrank	1.0	-6.082	9.67277	3.824	0.0506
Gehan-Wilcoxon	n_j	-246.0	21523.4	2.812	0.0936
PPP Wilcoxon	$\hat{S}(t_{(j)})$	-3.447	3.90095	3.046	0.0810

Since the data appear to satisfy the proportional hazards assumption to a greater degree than the proportional survival odds assumption, the logrank test yields a higher value for the test and a smaller p -value than does either the Gehan test or the Peto-Peto-Prentice (PPP) Wilcoxon test. Of course, one must prespecify which of these tests will be used a priori or the size of the test will be inflated.

An alternative is to use Gastwirth's maximin combination of the logrank and PPP Wilcoxon tests. The covariance between the two test statistics is 5.63501, which yields a correlation of 0.91735. Then, using the expression presented in (4.125), the maximin-efficient robust test (MERT) test value is $X_m^2 = 3.572$ with $p \leq 0.059$.

Although not statistically significant, these tests provide a test of differences between the complete survival curves of the two groups, rather than a test of the difference between groups at a particular point in time, as in Example 9.1.

Based on the logrank test, the Peto, Pike et al. (1976) estimate of the log hazard ratio or relative risk is $\hat{\beta}_{RR} = -0.629$ with estimated variance $\hat{V}(\hat{\beta}_{RR}) = 0.1034$, which yields an estimate of the relative risk of 0.533 with asymmetric 95% confidence limits (0.284, 1.001). A Cox proportional hazards regression model fit to these data with a single covariate for treatment effect yields a regression coefficient of -0.677 with $S.E. = 0.364$, and with an estimated relative risk of 0.508 and confidence limits of (0.249, 1.036), close to those provided by the score-based estimate. This quantity corresponds to the effect tested by the logrank test.

Likewise, for the PPP Wilcoxon test, the score-based estimate of the log cumulative event odds ratio is $\hat{\beta}_{OR} = -0.884$ with estimated variance $\hat{V}(\hat{\beta}_{OR}) = 0.2564$, which yields an estimate of the cumulative event odds ratio of 0.413 with asymmetric 95%

confidence limits (0.153, 1.115). The corresponding estimate of the survival odds ratio is $1/0.413 = 2.42$ with 95% confidence limits (0.897, 6.53). This quantity corresponds to the effect tested by the PPP Wilcoxon test.

Example 9.5 *Nephropathy in the DCCT (continued)*

For the analysis of the cumulative incidence of developing microalbuminuria in the secondary intervention cohort of the DCCT presented in Example 9.2, the study protocol specified a priori that the Mantel-logrank test would be used in the primary analysis to assess the difference between groups. This test yields $X^2 = 9.126$ with $p \leq 0.0026$. For comparison, the Peto-Peto-Prentice Wilcoxon test yields $X^2 = 8.452$ with $p \leq 0.0037$. For this example, the two tests are nearly equivalent.

The relative risk estimated from a Cox proportional hazards model is 1.62 with 95% confidence limits of (1.18, 2.22). This corresponds to a 38% reduction in the risk of progression to microalbuminuria among patients in the intensive versus conventional treatment groups. An additional calculation that adjusts for the log of the albumin excretion rate at baseline yields a 42.5% reduction in risk with intensive therapy.

9.3.6 SAS PROC LIFETEST: Tests of Significance

LIFETEST also provides a *strata* statement that allows for a stratified analysis by one or more categorical variables, and the option to conduct a variety of tests, whether stratified or not. For the squamous cell carcinoma data set of Example 9.1 with continuous (or nearly so) survival times, the following statements would provide the logrank (*Log-Rank*), Gehan-Wilcoxon (*Wilcoxon*), and PPP Wilcoxon (*Peto*) test in an unstratified analysis:

```
proc lifetest alpha=0.05 notable;
  time time*delta(0);
  strata / group=treatmnt test = logrank wilcoxon peto;
```

with the results

Test of Equality over Group			
Test	Chi-Square	DF	Pr >
Log-Rank	3.8245	1	0.0505
Wilcoxon	2.8116	1	0.0936
Peto	3.1073	1	0.0779

However, the Peto test value differs slightly from that presented in Example 9.4 owing to the fact that a slightly different estimate of the survival function is employed using $n_j + 1$ in the denominator to compute the event probabilities in (9.17).

PROC LIFETEST also provides a *test* statement that conducts either a logrank or a Gehan-Wilcoxon scores linear rank test that can be used to assess the association between a quantitative covariate and the event times. These tests are approximately

equivalent to the value of the Wald test of the covariate when used in either a proportional hazards or a proportional odds model, respectively. These tests could also be used with a binary covariate representing the treatment groups, as in the following statements:

```
proc lifetest alpha=0.05 notable;
  time time*delta(0); test treatmnt;
```

The tests, however, are based on the asymptotic permutational variance of the linear rank test and not the hypergeometric variance, as employed in the family of weighted Mantel-Haenszel tests. For the squamous cell carcinoma data, the logrank scores chi-square test value for the effect of treatment group is 3.579 with $p \leq 0.0585$, and the Wilcoxon scores test value is 3.120 with $p \leq 0.0773$. These tests do not correspond to the members of the G^{ρ} family of weighted Mantel-Haenszel tests, which, in general, are preferred.

9.4 PROPORTIONAL HAZARDS MODELS

The logistic and Poisson regression models of previous chapters assess the effects of covariates on the risk of the occurrence of an outcome or event, or multiple events, respectively, over a fixed period of time without consideration of the exact times at which events occurred, or the precise interval of time during which the event(s) occurred. The Cox (1972) proportional hazards (PH) regression model provides for the assessment of covariate effects on the risk of an event over time, where some of the subjects may not have experienced the event during the period of study, or may have a censored event time. The most common example is the time to death and the time of exposure (censoring) if a subject is still alive at the time that follow-up closed. A further generalization, the multiplicative intensity model of Aalen (1978) and Andersen and Gill (1992), allows the assessment of covariate effects on risk of multiple or recurrent events during the period of exposure. These and other generalizations are described later in this chapter.

9.4.1 Cox's Proportional Hazards Model

The assumption of proportional hazards is defined such that

$$\lambda(t) = \phi \lambda_0(t) \quad \text{and} \quad S(t) = S_0(t)^{\phi} \quad (9.53)$$

for some real constant of proportionality ϕ . Because the hazard function is the instantaneous risk of the event over time, the hazard ratio ϕ is a measure of relative risk. Now let \mathbf{X} be a vector of covariates measured at baseline ($t = 0$). Later we generalize to allow for time-varying (time-dependent) covariates. A proportional hazards regression model is one where

$$\phi = h(\mathbf{x}, \boldsymbol{\beta}) \quad (9.54)$$

for some smooth function $h(\mathbf{x}, \boldsymbol{\beta})$. Since the constant of proportionality must be positive, it is convenient to adopt a *multiplicative risk model*,

$$h(\mathbf{x}, \boldsymbol{\beta}) = e^{\mathbf{x}' \boldsymbol{\beta}} = e^{\sum_{j=1}^p x_j \beta_j} = \prod_{j=1}^p e^{x_j \beta_j}, \quad (9.55)$$

where the covariate effects on risk (λ) are multiplicative. The coefficient for the j th covariate is such that

$$\beta_j = \log(\phi) \text{ per } [\Delta X_j = 1] = \Delta \log[\lambda(t)] \text{ per } [\Delta X_j = 1], \quad (9.56)$$

and e^{β_j} is the log relative hazard (relative risk) per unit change in the covariate.

Such regression models can be derived parametrically, by assuming a specified form for $\lambda_0(t)$ and thus $\lambda(t|\mathbf{x})$ (see Section 9.7.3). To avoid the need to specify the shape of the hazard function, or a specific distribution, Cox (1972) proposed a *semiparametric* proportional hazards model with multiplicative risks such that

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t) e^{\mathbf{x}'_i \boldsymbol{\beta}} \quad \text{and} \quad \phi = e^{\mathbf{x}'_i \boldsymbol{\beta}} \quad \forall t, \quad (9.57)$$

where $\lambda_0(t)$ is an arbitrary background or nuisance hazard function interpreted as the hazard for an individual with covariate vector $\mathbf{x} = \mathbf{0}$. In this case the cumulative hazard is

$$\Lambda(t|\mathbf{x}_i) = \Lambda_0(t) e^{\mathbf{x}'_i \boldsymbol{\beta}} = \int_0^{t_i} \lambda_0(u) e^{\mathbf{x}'_i \boldsymbol{\beta}} du \quad (9.58)$$

with corresponding survival function

$$S(t|\mathbf{x}_i) = S_0(t)^{e^{\mathbf{x}'_i \boldsymbol{\beta}}}. \quad (9.59)$$

The background hazard function $\lambda_0(t)$ determines the shape of the background survival function $S_0(t)$ that is then expanded or shrunk through the effects of the covariates.

Assume that the observations are observed in continuous time with no tied event times. Then let $t_{(1)} < t_{(2)} < \dots < t_{(d)}$ refer to the d unique event times. To derive estimates of the coefficients, Cox (1972) employed a *partial likelihood* motivated heuristically as follows.

Let $\mathbf{x}_{(j)}$ be the covariate vector for the patient who experiences the event at the j th event time $t_{(j)}$. Then the probability of that patient experiencing the event at time $t_{(j)}$, given the patient is still at risk, is

$$\lambda(t_{(j)}|\mathbf{x}_{(j)}) = \lambda_0(t_{(j)}) e^{\mathbf{x}'_{(j)} \boldsymbol{\beta}}. \quad (9.60)$$

Let $R(t_{(j)})$ denote the *risk set*

$$R(t_{(j)}) = \{\ell : t_\ell \geq t_{(j)}\}, \quad (9.61)$$

consisting of the indices of all subjects in the original cohort still at risk at time $t_{(j)}$. Then the total probability of an event occurring at $t_{(j)}$ among all subjects in the risk set is

$$\sum_{\ell \in R(t_{(j)})} \lambda(t_{(j)} | \mathbf{x}_\ell) = \sum_{\ell \in R(t_{(j)})} \lambda_0(t_{(j)}) e^{\mathbf{x}'_\ell \boldsymbol{\beta}}. \quad (9.62)$$

Therefore, the conditional probability of an event occurring at $t_{(j)}$ in a subject with covariate vector $\mathbf{x}_{(j)}$ at that time is

$$\frac{\lambda(t_{(j)} | \mathbf{x}_{(j)})}{\sum_{\ell \in R(t_{(j)})} \lambda(t_{(j)} | \mathbf{x}_\ell)} = \frac{e^{\mathbf{x}'_{(j)} \boldsymbol{\beta}}}{\sum_{\ell \in R(t_{(j)})} e^{\mathbf{x}'_\ell \boldsymbol{\beta}}}. \quad (9.63)$$

This leads to the likelihood

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{j=1}^{d_*} \frac{e^{\mathbf{x}'_{(j)} \boldsymbol{\beta}}}{\sum_{\ell \in R(t_{(j)})} e^{\mathbf{x}'_\ell \boldsymbol{\beta}}}, \quad (9.64)$$

that is a function only of the times at which events occur. In fact, however, as is shown below, this is not a complete likelihood but rather is a partial likelihood, the complete likelihood involving other terms. Thus, the partial likelihood is designated as $\tilde{L}(\boldsymbol{\beta})$.

This likelihood can also be expressed in terms of individuals rather than event times. Using the joint observations (δ_i, t_i) , where δ_i is the indicator variable that denotes either the event or right censoring at time t_i , $i = 1, \dots, N$, Cox's partial likelihood is

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{i=1}^N \left[\frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\sum_{\ell \in R(t_i)} e^{\mathbf{x}'_\ell \boldsymbol{\beta}}} \right]^{\delta_i}, \quad (9.65)$$

where only those individuals with an event ($\delta_i = 1$) contribute to the likelihood. The partial likelihood can also be expressed as

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{i=1}^N \left[\frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\sum_{\ell=1}^N Y_\ell(t_i) e^{\mathbf{x}'_\ell \boldsymbol{\beta}}} \right]^{\delta_i} = \frac{\exp \left[\sum_{i=1}^N \delta_i \mathbf{x}'_i \boldsymbol{\beta} \right]}{\prod_{i=1}^N \left[\sum_{\ell=1}^N Y_\ell(t_i) e^{\mathbf{x}'_\ell \boldsymbol{\beta}} \right]^{\delta_i}}, \quad (9.66)$$

where $Y(t)$ refers to a time-dependent at-risk indicator variable to denote being at risk at time t . In the analysis of survival times, or the time of a single event,

$$Y_i(u) = \begin{cases} 1 : u \leq t_i \\ 0 : u > t_i \end{cases}, \quad u \geq 0. \quad (9.67)$$

To see why Cox's likelihood is a partial likelihood, it is useful to examine the full model and its partition into two partial likelihoods as described by Johansen (1983). From (9.53) and (9.57), the full likelihood under random censoring in (9.6) is

$$\begin{aligned} L[\beta, \lambda_0(t)] &= \prod_{i=1}^N \lambda(t_i | \mathbf{x}_i)^{\delta_i} S(t_i | \mathbf{x}_i) \\ &= \prod_{i=1}^N \left[\lambda_0(t_i) e^{\mathbf{x}'_i \beta} \right]^{\delta_i} \left[\exp \left(- \int_0^{t_i} \lambda_0(u) e^{\mathbf{x}'_i \beta} du \right) \right] \\ &= \prod_{i=1}^N \left[\lambda_0(X_i) e^{\mathbf{x}'_i \beta} \right]^{\delta_i} \prod_{i=1}^N \left[\exp \left(- \int_0^{\infty} Y_i(u) \lambda_0(u) e^{\mathbf{x}'_i \beta} du \right) \right] \end{aligned} \quad (9.68)$$

in terms of the at-risk indicator process $Y(u)$ for each subject. Then let

$$B(u) = \sum_{\ell=1}^N Y_{\ell}(u) e^{\mathbf{x}'_{\ell} \beta}, \quad u \geq 0. \quad (9.69)$$

Multiplying and dividing by $B(u)$, and rearranging terms, the likelihood becomes

$$\begin{aligned} L[\beta, \lambda_0(t)] &= \prod_{i=1}^N \left[\frac{e^{\mathbf{x}'_i \beta}}{B(t_i)} \right]^{\delta_i} \prod_{i=1}^N [B(t_i) \lambda_0(t_i)]^{\delta_i} \exp \left[- \int_0^{\infty} B(u) \lambda_0(u) du \right] \\ &= \tilde{L}_{(1)}(\beta) \tilde{L}_{(2)}[\beta, \lambda_0(t)], \end{aligned} \quad (9.70)$$

where $\tilde{L}_1(\beta)$ is Cox's partial likelihood $\tilde{L}(\beta)$ in (9.66) and $\tilde{L}_{(2)}[\beta, \lambda_0(t)]$ is the remaining partial likelihood involving both the coefficients β and the baseline hazard function $\lambda_0(t)$.

The technical difficulty is to demonstrate that a solution to the estimating equation based on the partial likelihood $\tilde{L}_{(1)}(\beta)$ provides estimates with the same properties as those of an *MLE* that maximizes a full likelihood. Cox (1972) simply applied the theory of maximum likelihood to the partial likelihood to provide estimates of the coefficients and to describe their asymptotic distribution, as though the associated score vector and information matrix were obtained from a full likelihood. However, this was received with some skepticism. Kalbfleisch and Prentice (1973) showed that Cox's partial likelihood could be obtained as a marginal likelihood when there were no tied observations. Cox (1975) provided a justification for the validity of inferences using a conditional likelihood argument. Tsiatis (1981) and others, under specific assumptions, showed that the partial likelihood score vector is asymptotically normally distributed with mean zero and covariance matrix equal to the partial likelihood information function, and thus that the partial maximum likelihood estimates were also asymptotically normally distributed like those based on a full model. A rigorous justification was also provided using counting process methods by Andersen and Gill (1982) in the context of the more general multiplicative intensity model that generalizes the Cox model to recurrent events; see also Gill (1984).

There are many generalizations of the basic proportional hazards model. A few of the most important are now described.

9.4.2 Stratified Models

Assume that the sample of N observations is divided among K mutually exclusive strata, and that there is a stratum-specific background hazard within each stratum. If we also assume that the covariate effects on the relative risks are homogeneous over strata, then for the h th stratum the stratified PH model specifies that

$$\lambda_h(t|\mathbf{x}) = \lambda_{0h}(t) e^{\mathbf{x}'\boldsymbol{\beta}_h}, \quad h = 1, \dots, K, \quad (9.71)$$

where $\lambda_{0h}(t)$ may vary over strata. For example, if we wished to assess the effects of age and weight on the risk of death, it might be appropriate to consider a model stratified by gender, where the background hazards would be allowed to differ for men and women, but where the effects of age and weight on relative risks are assumed to be the same for men and women.

A more general model arises when the covariate effects are also assumed to be heterogeneous over strata, in which case the model specifies that

$$\lambda_h(t|\mathbf{x}) = \lambda_{0h}(t) e^{\mathbf{x}'\boldsymbol{\beta}_h}, \quad h = 1, \dots, K, \quad (9.72)$$

where the $\boldsymbol{\beta}_h$ also differ among strata. Thus, the background hazards differ among strata as well as the constant of proportionality for each covariate. In this more general case, the stratified PH model partial likelihood is

$$\tilde{L}_i(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K) = \prod_{h=1}^K \prod_{i=1}^{N_h} \left[\frac{e^{\mathbf{x}'_{i(h)}\boldsymbol{\beta}_h}}{\sum_{\ell \in R_h(t_{i(h)})} e^{\mathbf{x}'_{\ell(h)}\boldsymbol{\beta}_h}} \right]^{\delta_{i(h)}}, \quad (9.73)$$

where $i(h)$ denotes the i th subject within the h th stratum and where the risk set is defined only among the N_h observations in the h th stratum: $R_h(t_{i(h)}) = \{\ell \in \text{stratum } h : t_{\ell(h)} \geq t_{i(h)}\}$.

This stratified model implies that there is an interaction between the stratification variable and the background hazard function and with the other covariates that enter into the model through $\mathbf{x}'\boldsymbol{\beta}_h$. When the coefficients are homogeneous over stratum as in (9.71), it is assumed that $\boldsymbol{\beta}_h = \boldsymbol{\beta}$ for all strata, which yields a slight simplification of the partial likelihood. In this case, there is an interaction between stratum and the background hazard only. Thall and Lachin (1986) describe these and other forms of covariate interactions in the PH model.

Stratified models are also useful when the proportional hazards assumption does not apply in the aggregate population, but it does apply within strata. For example, when the relative risks differ within intervals of time, so that the proportional hazards assumption does not apply in general, it may still apply within separate intervals of time, such as $0 < t \leq 1$, $1 < t \leq 2$, and so forth. In this case a stratified model as in (9.73) could allow for proportional hazards within time strata, with differences in covariate effects on risk among the different time strata.

9.4.3 Time-Dependent Covariates

In some instances the covariate process may vary over time, in which case the covariate is termed *time-dependent*. Since the hazard function is the instantaneous probability of the event at time t given exposure to time t , or being at risk at time t , any aspect of the history of the covariate process up to time t may be incorporated into the model to assess the effect of the covariate on the relative risk of the event over time. Technically, a time-dependent covariate is assumed to be observable up to, but not including, t itself, or $X(t) \in \mathfrak{X}_{t^-}$, which represents the history of the process up to the instant before time t , designated as t^- . Then the time-dependent PH model is of the form

$$\lambda[t|\mathbf{x}(t)] = \lambda_0(t) e^{\mathbf{x}(t)'\boldsymbol{\beta}}, \quad (9.74)$$

where the j th coefficient now represents

$$\beta_j = \log(\phi) \text{ per } [\Delta X_j(t) = 1] = \Delta \log \{\lambda[t|X_j(t)]\} \text{ per } [\Delta X_j = 1]. \quad (9.75)$$

For a fixed time (not time-dependent) covariate, the value over time is the same as that at baseline as employed in (9.64). With a time-dependent covariate vector, the basic partial likelihood then becomes

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{i=1}^N \left[\frac{e^{\mathbf{x}_i(t_i)'\boldsymbol{\beta}}}{\sum_{\ell \in R(t_i)} e^{\mathbf{x}_\ell(t_i)'\boldsymbol{\beta}}} \right]^{\delta_i}. \quad (9.76)$$

To fit this model then requires that the time-dependent covariates be evaluated for all subjects at risk at each event time.

9.4.4 Fitting the Model

Consider the unstratified model with a possibly time-dependent covariate vector with no tied event times for which the partial likelihood is (9.76). Then the j th element of the score vector $\mathbf{U}(\boldsymbol{\beta})$ corresponding to coefficient β_j is

$$U(\boldsymbol{\beta})_{\beta_j} = \frac{\partial \log \tilde{L}(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^N \delta_i [x_{ij}(t_i) - \bar{x}_j(t_i, \boldsymbol{\beta})], \quad (9.77)$$

where $E[x_{ij}(t_i) | \boldsymbol{\beta}] = \bar{x}_j(t_i, \boldsymbol{\beta})$ and

$$\bar{x}_j(t_i, \boldsymbol{\beta}) = \frac{\sum_{\ell \in R(t_i)} x_{\ell j}(t_i) e^{\mathbf{x}_\ell(t_i)'\boldsymbol{\beta}}}{\sum_{\ell \in R(t_i)} e^{\mathbf{x}_\ell(t_i)'\boldsymbol{\beta}}} = \frac{\sum_{\ell=1}^N Y_\ell(t_i) x_{\ell j}(t_i) e^{\mathbf{x}_\ell(t_i)'\boldsymbol{\beta}}}{\sum_{\ell=1}^N Y_\ell(t_i) e^{\mathbf{x}_\ell(t_i)'\boldsymbol{\beta}}} \quad (9.78)$$

is the weighted average of the covariate among all those at risk at time t_i . Thus, the score equation is of the form $\sum(\text{observed} - \text{expected})$ expressed in terms of

the covariate values for the i th subject who experiences the event at time t_i . The contribution of each subject to the total score is also known as the Schoenfeld (1982) or score residual.

The corresponding information matrix, where $\mathbf{i}(\boldsymbol{\beta}) = \mathbf{I}(\boldsymbol{\beta})$, then has elements

$$\mathbf{I}(\boldsymbol{\beta})_{jk} = E \left[\frac{-\partial \log \tilde{L}(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_h} \right] = \sum_{i=1}^N \delta_i [C_{ijk}(t_i, \boldsymbol{\beta}) - \bar{x}_j(t_i, \boldsymbol{\beta}) \bar{x}_k(t_i, \boldsymbol{\beta})], \quad (9.79)$$

with

$$\begin{aligned} C_{ijk}(t_i, \boldsymbol{\beta}) &= \frac{\sum_{\ell \in R(t_i)} x_{\ell j}(t_i) x_{\ell k}(t_i) e^{\mathbf{x}_{\ell}(t_i)' \boldsymbol{\beta}}}{\sum_{\ell \in R(t_i)} e^{\mathbf{x}_{\ell}'(t_i) \boldsymbol{\beta}}} \\ &= \frac{\sum_{\ell=1}^N Y_{\ell}(t_i) x_{\ell j}(t_i) x_{\ell k}(t_i) e^{\mathbf{x}_{\ell}(t_i)' \boldsymbol{\beta}}}{\sum_{\ell=1}^N Y_{\ell}(t_i) e^{\mathbf{x}_{\ell}(t_i)' \boldsymbol{\beta}}} \end{aligned} \quad (9.80)$$

for $1 \leq j \leq k \leq p$. The model is then fit using the Newton-Raphson iteration.

In a stratified model with homogeneous covariate effects among strata ($\beta_h = \boldsymbol{\beta} \forall h$) as in (9.71), the score equation for the j th coefficient is simply $U(\boldsymbol{\beta})_{\beta_j} = \sum_{h=1}^K U_h(\boldsymbol{\beta})_{\beta_j}$, where the $U_h(\boldsymbol{\beta})_{\beta_j}$ is the score equation for the j th covariate evaluated with respect to those subjects within the h th stratum only, $h = 1, \dots, K$. Likewise, using summations over subjects within strata, the covariate means within the h th strata $\{\bar{x}_{j(h)}(t_{i(h)}, \boldsymbol{\beta})\}$ are computed with respect to the risk set comprising subjects within that stratum $R_h(t_{i(h)})$. The information matrix then has elements $\mathbf{I}(\boldsymbol{\beta})_{jk} = \sum_{h=1}^K \mathbf{I}_h(\boldsymbol{\beta})_{jk}$, where $\mathbf{I}_h(\boldsymbol{\beta})$ is computed from the subjects in the h th stratum only.

In a stratified model with stratum-specific covariate effects β_h for $h = 1, \dots, K$ as in (9.72), the score vector contains Kp elements,

$$\mathbf{U}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K) = \left[\mathbf{U}_1(\boldsymbol{\beta}_1)^T \parallel \dots \parallel \mathbf{U}_K(\boldsymbol{\beta}_K)^T \right]^T, \quad (9.81)$$

where $\mathbf{U}_h(\boldsymbol{\beta}_h)$ is computed as in (9.77) among the subjects within the h th stratum. Since the strata are independent, the information matrix is a $Kp \times Kp$ block diagonal matrix of the form

$$\mathbf{I}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K) = \text{blockdiag} [\mathbf{I}_1(\boldsymbol{\beta}_1) \cdots \mathbf{I}_K(\boldsymbol{\beta}_K)]. \quad (9.82)$$

The information matrix for the h th stratum $\mathbf{I}_h(\boldsymbol{\beta}_h)$ has elements $\mathbf{I}_h(\boldsymbol{\beta}_h)_{jk}$ evaluated with respect to the subjects within that stratum.

It is also straightforward to further generalize the above to the case with stratum-specific effects for some covariates and homogeneous effects for others.

Although the model is based on a partial likelihood, the three general families of tests can be employed as for any other likelihood-based model: the Wald test, efficient scores test, or the likelihood ratio test. As a problem it is readily shown that the Mantel-logrank test is the efficient scores test for the treatment group effect in the Cox model using the Peto-Breslow adjustment for ties (see Section 9.4.6.4).

9.4.5 Robust Inference

Inferences using the Cox model are dependent on the proportional hazards assumption that may not apply. In this case, one may still be interested in the Cox model coefficients as estimates of log *average* hazard ratios over time, even though a constant hazards ratio does not apply. The model-based standard errors and tests of significance in this case are possibly biased because of the departure from the proportional hazards assumption.

Gail et al. (1984) and Lagakos (1988), among others, have also shown that the estimates of the model coefficients in a Cox model are biased when important covariates are omitted from the model, even when the covariates are independent of each other. Whereas the coefficients in a simple exponential model are not biased by the omission of an independent relevant covariate, as shown in Problem 7.6.3, those in the Cox model are biased because of the presence of censoring, and because the estimates are obtained from a partial likelihood rather than the full likelihood. Thus, model-based inferences are also biased when an important covariate is omitted from the Cox model.

The Cox model also assumes that the same background hazard function applies to all subjects in the population. A generalization of this assumption leads to a *frailty model*, in which it is assumed that the background hazard function has some distribution between subjects in the population (Clayton and Cuzick, 1985; see also Andersen et al., 1993). This is similar to a regression model such as logistic or Poisson regression, where it is assumed that the intercept in the model is distributed randomly in the population. Whereas the robust sandwich estimator will account for overdispersion in a logistic or Poisson regression model, it does not account for frailty in the Cox model because the coefficients are distinct from those in the marginal Cox model.

In cases of model misspecification (other than frailty), proper estimates of the covariance matrix of the coefficient estimates, and related confidence intervals and statistical tests can be provided by the robust information sandwich (see Section A.9). The score vector for the i th subject is

$$U_i(\beta) = \frac{\partial \log \tilde{L}_i(\beta)}{\partial \beta} = \delta_i [\mathbf{x}_i(t_i) - \bar{\mathbf{x}}(t_i, \beta)], \quad (9.83)$$

where

$$\bar{\mathbf{x}}(t_i, \beta) = \frac{\sum_{j=1}^N Y_j(t_i) \mathbf{x}_j(t_i) e^{\mathbf{x}_j(t_i)' \beta}}{\sum_{j=1}^N Y_j(t_i) e^{\mathbf{x}_j(t_i)' \beta}} = \frac{S_1(\beta, t_i)}{S_0(\beta, t_i)} = E[\mathbf{x}(t_i) | \beta]. \quad (9.84)$$

However, because the proportional hazards model is based on a partial rather than a full likelihood, the score vectors are not independently and identically distributed with expectation zero. Thus, Lin and Wei (1989) describe a modification of the information sandwich that provides a consistent estimate of the covariance matrix of the coefficient estimates under model misspecification; see also Lin (1994). They

show that

$$\begin{aligned}\mathbf{W}_i(\boldsymbol{\beta}) &= \mathbf{U}_i(\boldsymbol{\beta}) - E[\mathbf{U}_i(\boldsymbol{\beta})] \\ E[\mathbf{U}_i(\boldsymbol{\beta})] &= \sum_{\ell=1}^N \frac{\delta_\ell Y_i(t_\ell) e^{\mathbf{x}_i(t_\ell)' \boldsymbol{\beta}}}{S_0(\boldsymbol{\beta}, t_\ell)} (\mathbf{x}_i(t_\ell) - \bar{\mathbf{x}}(t_\ell, \boldsymbol{\beta})),\end{aligned}\quad (9.85)$$

where $E[\mathbf{U}_i(\boldsymbol{\beta})]$ is a weighted average of the scores of all subjects with events prior to time t_i with respect to the covariate value for the i th subject at that time. Then, the $\{\mathbf{W}_i(\boldsymbol{\beta})\}$ are *i.i.d.*

Thus, the robust estimator of the covariance matrix of the score vector is

$$\widehat{V}[\mathbf{U}(\boldsymbol{\beta})] = \widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^N \mathbf{W}_i(\widehat{\boldsymbol{\beta}}) \mathbf{W}_i(\widehat{\boldsymbol{\beta}})' \quad (9.86)$$

and the robust estimate of the covariance matrix of the coefficient vector is

$$\widehat{\boldsymbol{\Sigma}}_R(\widehat{\boldsymbol{\beta}}) = \mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1} \widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}}) \mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1}. \quad (9.87)$$

This provides robust confidence limits for the parameter estimates and robust Wald and score tests for the parameters. These computations are provided by recent versions of SAS PROC PHREG. By simulation, Lin and Wei (1989) show that the model-based Wald and score tests may have markedly inflated size when important covariates have inadvertently been omitted from the model, whereas the robust estimates retain their nominal size.

A robust model score test of $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \mathbf{0}$ may be computed by evaluating the score vector $\mathbf{U}_i(\boldsymbol{\beta}_0)$, the centered scores $\mathbf{W}_i(\boldsymbol{\beta}_0)$, and the matrix $\mathbf{J}(\boldsymbol{\beta}_0)$ under the null hypothesis. From (9.85) the centered score vector for the i th subject evaluated under H_0 is

$$\mathbf{W}_i(\boldsymbol{\beta}_0) = \delta_i [\mathbf{x}_i(t_i) - \bar{\mathbf{x}}(t_i)] - \sum_{\ell=1}^N \frac{\delta_\ell Y_i(t_\ell)}{n(t_\ell)} (\mathbf{x}_i(t_\ell) - \bar{\mathbf{x}}(t_\ell)), \quad (9.88)$$

where $\bar{\mathbf{x}}(t)$ is the unweighted mean vector of covariate values among the $n(t)$ subjects at risk at time t . The robust covariance matrix under H_0 is then obtained as in (9.86) by $\widehat{\mathbf{J}}(\boldsymbol{\beta}_0) = \sum_{i=1}^N \mathbf{W}_i(\boldsymbol{\beta}_0) \mathbf{W}_i(\boldsymbol{\beta}_0)'$. Using the total score vector $\mathbf{W}(\boldsymbol{\beta}_0) = \sum_i \mathbf{W}_i(\boldsymbol{\beta}_0)$, the robust model score test is then provided by $X^2 = \mathbf{W}(\boldsymbol{\beta}_0)' \widehat{\mathbf{J}}(\boldsymbol{\beta}_0)^{-1} \mathbf{W}(\boldsymbol{\beta}_0)$, which asymptotically is distributed as chi-square on p *df* under H_0 . Similarly, robust score tests may be obtained for the individual parameters of the model, although the computations are more tedious (see Section A.7.3).

9.4.6 Adjustments for Tied Observations

The above presentation of the Cox PH model assumes that no two subjects experience the event at the same time, or that there are no tied event times, such as where multiple subjects die on the same study day. One crude and unsatisfactory option in this instance is to arbitrarily (by chance) break the ties and order them. With one

or two ties, and a large sample size, this will have a trivial effect on the analysis. However, when there are many ties, or ties arise because of the manner in which the observations are recorded, the analysis should include an appropriate adjustment for ties.

9.4.6.1 Discrete and Grouped Failure Time Data The simplest structure for a model that allows tied event times is to assume that events can only occur at fixed times $\tau_1 < \tau_2 < \dots < \tau_K$. However, this is unrealistic because almost always events occur continuously over time but may only be *observed* within grouped intervals of time, where the j th interval includes all times $A_j = (\tau_{j-1}, \tau_j]$. For example, in many studies, an examination or procedure must be performed to determine whether an event has occurred, in which case events during the j th interval will only be observed to have occurred on the examination conducted at τ_j . This structure assumes that *fixed intervals* apply to all subjects, or that all subjects are examined at the fixed times $\{\tau_j\}$. When the observation times vary from subject to subject, then the observations are interval censored, in which case different models will apply (see below). In some cases, however, a fixed interval model will apply approximately to interval-censored data.

Prentice and Gloeckler (1978) describe the generalization of the proportional hazards model to such discrete or grouped time data. The continuous-time proportional hazards model specifies that the survival probabilities satisfy the relationship $S(\tau_j | \mathbf{x}_i) = S_0(\tau_j)^{\exp(\mathbf{x}'_i \boldsymbol{\beta})}$ at any time τ_j , where $S_0(\tau_j)$ is the background survival function for a subject with covariate vector $\mathbf{x} = \mathbf{0}$. In the grouped-time model,

$$S_0(\tau_j) = \exp \left[- \int_0^{\tau_j} \lambda_0(s) \, ds \right], \quad (9.89)$$

however, no information is provided regarding the form of the underlying hazard function $\lambda_0(\tau_j)$. In either the grouped- or discrete-time model, the background conditional probability of a subject with covariate vector $\mathbf{x} = \mathbf{0}$ surviving interval A_j is

$$\varphi_{0j} = \frac{S_0(\tau_j)}{S_0(\tau_{j-1})} = \exp \left[- \int_{\tau_{j-1}}^{\tau_j} \lambda_0(s) \, ds \right], \quad (9.90)$$

and the conditional probability that the event is observed at τ_j is $1 - \varphi_{0j}$. Note that φ_{0j} is analogous to the continuation probability $1 - \pi_j$ in (9.9) in the Kaplan-Meier construction. Under the proportional hazards model, it follows that $\varphi_{j|\mathbf{x}} = \varphi_{0j}^{\exp(\mathbf{x}' \boldsymbol{\beta})}$.

Let a_i denote the final interval during which the i th subject was observed to be at risk, $a_i \in (1, \dots, K + 1)$, where the last possible interval is $A_{K+1} = (\tau_K, \infty)$; and let δ_i be the indicator variable to denote event ($\delta_i = 1$) or censoring ($\delta_i = 0$). Any subject censored during the j th interval ($a_i = j, \delta_i = 0$) is assumed not to be at risk during that interval. Any observations event free and under follow-up at the end of the study are right censored after τ_K . Thus, only censored observations have values $a_i = K + 1$. From (9.6) the likelihood function for a sample of N observations under

a random censoring model is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^N \prod_{j=1}^{K+1} \left[\left(1 - \varphi_{0j}^{\exp(\mathbf{x}'_i \boldsymbol{\beta})} \right)^{\delta_i I_{\{a_i=j\}}} \right] \left[\left(\varphi_{0j}^{\exp(\mathbf{x}'_i \boldsymbol{\beta})} \right)^{I_{\{a_i < j\}}} \right] \quad (9.91) \\ &= \prod_{i=1}^N \left(1 - \varphi_{0(a_i)}^{\exp(\mathbf{x}'_i \boldsymbol{\beta})} \right)^{\delta_i} \prod_{j=1}^{a_i-1} \left(\varphi_{0j}^{\exp(\mathbf{x}'_i \boldsymbol{\beta})} \right) \end{aligned}$$

for $\boldsymbol{\theta} = (\varphi_{01} \ \cdots \ \varphi_{0K} \ \beta_1 \ \cdots \ \beta_p)^T$.

Because the $\{\varphi_{0j}\}$ are probabilities, Prentice and Gloeckler (1978) use the complementary log(-log) link such that

$$\gamma_j = \log[-\log(\varphi_{0j})], \quad \varphi_{0j} = e^{-e^{\gamma_j}} \quad (9.92)$$

for $1 < j < K$. Substitution into the above yields estimating equations for the parameters $\boldsymbol{\theta} = (\gamma_1 \ \cdots \ \gamma_K \ \beta_1 \ \cdots \ \beta_p)^T$.

Note that the estimation of the relative risk coefficients requires joint estimation of the nuisance parameters $(\gamma_1, \dots, \gamma_K)$. For finite samples, as the number of such parameters (intervals) increases, the bias of the coefficient estimates also increases (cf. Cox and Hinkley, 1974).

The resulting model is then fit by maximum likelihood estimation. Whitehead (1989) shows that this model can be fit using PROC GENMOD or a program for generalized linear models that allows for a binomial distribution with a complementary log(-log) link as described in Section 7.1.5. An example is also provided in the SAS (2008b) description of PROC LOGISTIC.

9.4.6.2 Cox's Adjustment for Ties In practice, even when the event-time distribution is continuous, ties may be caused by coarsening of the time measures, such as measuring time to the week or month. For such cases, adjustments to the Cox PH model are employed to allow for ties. When there are tied event times, Cox (1972) suggested a discrete logistic model conditional on covariates \mathbf{X} for some time interval of length dt of the form

$$\frac{\lambda(t|\mathbf{x}) dt}{1 - \lambda(t|\mathbf{x}) dt} = \frac{\lambda_0(t) dt}{1 - \lambda_0(t) dt} e^{\mathbf{x}' \boldsymbol{\beta}}. \quad (9.93)$$

Therefore, $\lim dt \downarrow 0$ yields the PH model for continuous time in (9.57) since $\lim_{dt \downarrow 0} [1 - \lambda(t) dt] \rightarrow 1.0$. In Problem 9.14.12 it is also shown that the PH model results from a grouped time logistic model, as above, where $K \rightarrow \infty$. If we let

$$\alpha = \log \left[\frac{\lambda_0(t) dt}{1 - \lambda_0(t) dt} \right], \quad (9.94)$$

then

$$\frac{\lambda(t|\mathbf{x}) dt}{1 - \lambda(t|\mathbf{x}) dt} = e^{\alpha + \mathbf{x}' \boldsymbol{\beta}}. \quad (9.95)$$

Therefore, we have a logistic model of the form

$$\lambda(t|\mathbf{x}) dt = \frac{e^{\alpha + \mathbf{x}'\beta}}{1 + e^{\alpha + \mathbf{x}'\beta}} = \psi(\mathbf{x}). \quad (9.96)$$

Now, consider a set of N small but finite intervals of time. During the j th interval assume that n_j subjects are followed (at risk), of whom $d_j = m_{1j}$ subjects experience the outcome event and $n_j - d_j = m_{2j}$ subjects survive the interval event-free, where m_{1j} and m_{2j} are the numbers with and without the event in the notation of matched sampling used in Section 7.6. Let \mathbf{x}_{jk} denote the covariate vector for the k th subject to have the event at time $t_{(j)}$. Then conditioning on m_{1j} events and m_{2j} nonevents during the j th interval, the conditional likelihood is

$$\begin{aligned} L(\beta)_{j|m_{1j}, n_j} &= \frac{\prod_{k=1}^{m_{1j}} \psi(\mathbf{x}_{jk}) \prod_{k=m_{1j}+1}^{n_j} [1 - \psi(\mathbf{x}_{jk})]}{\sum_{\ell=1}^{\binom{n_j}{m_{1j}}} \prod_{k(\ell)=1}^{m_{1j}} \psi(\mathbf{x}_{jk(\ell)}) \prod_{k(\ell)=m_{1j}+1}^{n_j} [1 - \psi(\mathbf{x}_{jk(\ell)})]} \\ &= \frac{\prod_{k=1}^{m_{1j}} e^{\mathbf{x}'_{jk}\beta}}{\sum_{\ell=1}^{\binom{n_j}{m_{1j}}} \prod_{k(\ell)=1}^{m_{1j}} e^{\mathbf{x}'_{jk(\ell)}\beta}}. \end{aligned} \quad (9.97)$$

Because the successive intervals are conditionally independent, the total conditional likelihood is

$$L_c(\beta) = \prod_{j=1}^N L(\beta)_{j|m_{1j}, n_j}. \quad (9.98)$$

This likelihood is equivalent to that of the conditional logistic regression model for matched sets of Section 7.6.1. The score equations and the expressions for the information function are presented in (7.121)–(7.123).

This model is appropriate when there is some natural grouping of the event times. Thus, this model is appropriate for the instance described in Example 9.2, where the occurrence of the event can only be ascertained at fixed times during the study. The Prentice-Gloeckler grouped time model could also be applied to such data.

9.4.6.3 Kalbfleisch-Prentice Marginal Model Kalbfleisch and Prentice (1973; see also Kalbfleisch and Prentice, 1980) showed that Cox's partial likelihood with no ties could also be expressed as a marginal likelihood obtained by integrating out the background hazard function. For tied observations, they provide an expression for the corresponding marginal likelihood that is somewhat more computationally intensive than Cox's logistic likelihood.

9.4.6.4 Peto-Breslow Adjustment for Ties A computationally simple approach was suggested by Peto (1972) and Breslow (1974) as an approximation to a

precise model allowing for ties. Breslow (1974) showed that this provides an approximation to the marginal likelihood of Kalbfleisch and Prentice (1973) adjusting for ties. Again, let $\{t_{(j)}\}$ denote the set of J distinct event times among the total of d_* patients who experience the event, $J < d_*$. The events are assumed to occur in continuous time, but some ties are observed because of rounding or grouping of the times into small intervals, such that the number of tied event times is small relative to the total number of events d_* . Then let d_j denote the number of events observed among those at risk at time $t_{(j)}$. Generalizing (9.64), the approximate likelihood is

$$\tilde{L}_{PB}(\beta) = \prod_{j=1}^J \frac{\exp\left(\sum_{k=1}^{d_j} \mathbf{x}'_{jk} \beta\right)}{\left(\sum_{\ell \in R(t_j)} e^{\mathbf{x}'_\ell \beta}\right)^{d_j}}, \quad (9.99)$$

where \mathbf{x}_{jk} denotes the covariate vector for the k th subject to experience the event at time $t_{(j)}$, $1 \leq k \leq d_j$; and where the d_j subjects with the event at $t_{(j)}$ are included in the risk set at that time. This likelihood can also be expressed as a product over individuals as in (9.65). Thus, equations (9.77)–(9.80) and the expressions for the robust information matrix in Section 9.4.5 also apply to the Peto-Breslow approximate likelihood with tied observations.

9.4.7 Survival Function Estimation

Breslow (1972, 1974) proposed a simple estimate of the background survivor function $S_0(t)$, and thus of the hazard function $\lambda_0(t)$, under the assumption that the latter is piecewise constant between event times; i.e., $\lambda_0(t) = \lambda_j$ for $t_{j-1} < t \leq t_j$ for the j th event time. Evaluating the resulting full likelihood as a function of the J parameters $(\lambda_1, \dots, \lambda_J)$ and the coefficient vector β , the resulting estimating equations for the coefficients were the same as those in the PH model, whereas the estimates of the hazard function and the cumulative hazard function at time t_j were provided by the following expressions

$$\begin{aligned} \hat{\lambda}_j &= \frac{d_j}{(t_j - t_{j-1}) \sum_{\ell \in R(t_j)} e^{\mathbf{x}'_\ell \hat{\beta}}} \\ \hat{\Lambda}(t_j) &= \int_0^{t_j} \lambda_0(u) du = \sum_{k=1}^j \frac{d_k}{\sum_{\ell \in R(t_k)} e^{\mathbf{x}'_\ell \hat{\beta}}}. \end{aligned} \quad (9.100)$$

The latter is of the form of a Nelson-Aalen estimator analogous to (9.27). Thus, an asymptotically equivalent estimator of $S_0(t)$ in the form of the Kaplan-Meier estimate is provided by

$$\hat{S}_0(t) = \prod_{j: t_j \leq t} (1 - p_j), \quad (9.101)$$

where $p_j = (t_j - t_{j-1}) \hat{\lambda}_j$ is an estimate of π_j , the conditional probability of an event at t_j given remaining at risk.

In the first edition of their classic text, Kalbfleisch and Prentice (1980) also proposed a model that provided probability mass only at the observed times of failure that likewise included terms (π_1, \dots, π_J) as above. To simplify notation, let $\alpha_j = 1 - \pi_j$. The resulting likelihood function is of the form

$$L(\pi_1, \dots, \pi_J, \beta) = \prod_{j=1}^J \prod_{k=1}^{d_j} \left(1 - \alpha_k^{e^{\mathbf{x}'_k \beta}}\right) \prod_{\ell \in [R(t_k) - D_k]} \alpha_k^{e^{\mathbf{x}'_\ell \beta}}, \quad (9.102)$$

where D_k denotes the indices of those subjects with the event at time t_k ; that is, $[R(t_k) - D_k]$ denotes those subjects in the risk set who did not have the event at that time. The model then requires joint estimation of the probabilities and the coefficients. If fit by maximum likelihood, the resulting $\hat{\beta}$ will not necessarily equal those from the PH model. Thus, in practice the $\{\pi_j\}$ are estimated iteratively based on the PH model estimates $\hat{\beta}$. The resulting estimated baseline survival function is obtained by substituting the $p_j = \hat{\pi}_j$ into (9.101). Simplifications and approximations can also be obtained (see Collett, 1994).

Using either estimate, the resulting survival function estimate for an individual with covariate vector \mathbf{x} is provided by

$$\hat{S}(t|\mathbf{x}) = \hat{S}_0(t)^{e^{\mathbf{x}' \hat{\beta}}}. \quad (9.103)$$

This also provides for the computation of a covariate-adjusted estimated survival function within groups of subjects, analogous to the LSMEAN-adjusted probability estimates in a logistic regression model described in Chapter 7. Suppose that x_1 is a binary (0, 1) variable to represent treatment group with coefficient $\hat{\beta}_1$ and covariates X_2, \dots, X_p are a set of adjusting covariates with mean values $\bar{x}_2, \dots, \bar{x}_p$ and coefficients $\hat{\beta}_2, \dots, \hat{\beta}_p$. The covariate-adjusted estimated survival functions in the two groups are obtained as

$$\begin{aligned} \text{Group 0: } \hat{S}(t|\mathbf{x}) &= \hat{S}_0(t)^{e^{\bar{x}_2 \hat{\beta}_2 + \dots + \bar{x}_p \hat{\beta}_p}} \\ \text{Group 1: } \hat{S}(t|\mathbf{x}) &= \hat{S}_0(t)^{e^{\hat{\beta}_1 + \bar{x}_2 \hat{\beta}_2 + \dots + \bar{x}_p \hat{\beta}_p}}. \end{aligned} \quad (9.104)$$

The variance of the estimate $\hat{S}(t|\mathbf{x})$ is tedious. See, for example, Kalbfleisch and Prentice (1980, 2002). Given the estimate $\hat{V}[\hat{S}(t|\mathbf{x})]$, the variance of suitable transformations, such as the log, are then readily obtained.

9.4.8 Model Assumptions

Consider a model with a single possibly time-dependent covariate $X(t)$. The principal assumption of the proportional hazards model is that the effect of $X(t)$ on the background hazard function over time is described by the constant of proportionality $\phi = e^{x(t)\beta}$. To test this assumption, Cox (1972) proposed a test of $H_0: \phi[t|x(t)] = e^{x(t)\beta} \forall t$ of a constant hazard ratio over time against the alternative $H_1: \phi[t|x(t)] = h(t) \neq e^{x(t)\beta}$ that the relative hazard is a monotonically increasing or decreasing function of time $h(t)$, such as where $h(t) = \log(t)$. This specific alternative implies

that the true model includes an interaction term between the covariate and $h(t)$ such as $x(t)\beta_1 + x(t)h(t)\beta_2$. Note that no term for $h(t)$ is required in the exponent because this term, if added, would be absorbed into the background hazard function $\lambda_0(t)$. A test of $H_0: \beta_2 = 0$ then provides a test of the PH assumption for that covariate in the model against the specific alternative that the hazard ratio $\phi[t|x(t)]$ is assumed to be log linear in $h(t)$.

However, this is a test of a specific mode of departure from the PH model assumption. Numerous authors have proposed alternative assessments of this assumption, many based on graphical assessments. Lin and Wei (1991) present a review of these methods. Among the simplest is the following. Assume that we wish to assess the proportionality assumption for covariate Z when added to the covariate vector \mathbf{X} , where $\phi(t|z, \mathbf{x}) = \exp(z\gamma + \mathbf{x}'\boldsymbol{\beta})$. Then using the complementary log(-log) transformation of the survival function yields

$$\log(-\log[S(t|z, \mathbf{x})]) = z\gamma + \mathbf{x}'\boldsymbol{\beta} + \log(-\log[S_0(t|z, \mathbf{x})]). \quad (9.105)$$

This implies that if Z is used to construct separate strata ($h = 1, \dots, K$), then for fixed values of the other covariates (\mathbf{x}),

$$\log(-\log[\hat{S}_h(t|\mathbf{x})]) = \mathbf{x}'\hat{\boldsymbol{\beta}} + \log(-\log[\hat{S}_{0h}(t|\mathbf{x})]). \quad (9.106)$$

Thus, if the hazards are indeed proportional for different values of Z , now represented by strata, then the background hazard functions within strata should be proportional. In this case, plots of the functions $\log(-\log[\hat{S}_{0h}(t|\mathbf{x})])$ versus t or $\log(t)$ should have a constant difference, approximately, over time.

Various types of residual diagnostics have also been proposed that may be used to identify influential observations and to detect departures from the proportional hazards assumption. A summary of these methods is provided by Fleming and Harrington (1991). Many of these methods are also inherently graphical.

Lin et al. (1993) propose tests of the model assumptions that can be used with the martingale residuals of Therneau et al. (1990). A somewhat simpler test of the PH model assumption was also described by Lin (1991). Without showing the details, the basic idea is related to the properties of efficiently weighted Mantel-Haenszel tests where a unit weight for the difference ($O - E$) at each event time is asymptotically efficient against a proportional hazards alternative. Thus, if a different set of weights yields a test with a significantly greater value than the test with unit weights, this is an indication that the proportional hazards assumption does not apply. Lin (1991), therefore, considers the difference between the weighted sum of the score vectors of the form

$$T = \sum_i [\mathbf{U}_i(\boldsymbol{\beta}) - w(t_i)\mathbf{U}_i(\boldsymbol{\beta})] \quad (9.107)$$

for some weight $w(t_i) \neq 1$. To test the proportional hazards assumption against a proportional odds assumption, the test would employ $w(t_i) = \hat{S}(t_i)$ as would be used in a Peto-Peto-Prentice Wilcoxon test. Lin (1991) then describes a test of significance for the PH assumption based on this statistic and provides a program for its computation. A SAS macro is also available.

9.4.9 Explained Variation

Since the PH model is semiparametric and is based on a partial likelihood, no simple, intuitively appealing measure of explained variation arises naturally. Schemper (1990) proposed a measure (his V_2), which is defined from the ratio of the weighted sum of squares of the deviations of the empirical survival function for each individual over time with respect to the Cox model fitted survival function under the null ($\beta = 0$) and alternative ($\beta = \hat{\beta}$) hypotheses. O'Quigley et al. (1999) show that this measure is a Korn-Simon-like measure of explained variation (see Section A.8) in terms of survival probabilities that are weighted by the increments in the empirical cumulative distribution function of the event times.

Schemper's measure, however, is bounded above by some constant less than 1.0 when the model fits perfectly. Also, it is based on the Cox model estimated survival function rather than the hazard function on which the model is based. Computation requires estimation of the background survival function $S_0(t)$, from which one obtains an estimate of the conditional survival function as $\hat{S}(t|\mathbf{x}) = \hat{S}_0(t)^{\exp[\mathbf{x}'\hat{\beta}]}$. Either the Breslow (1974) or Kalbfleisch and Prentice (1980) estimate of $\hat{S}_0(t)$ can be employed as described in Section 9.4.7.

Korn and Simon (1990) also defined a measure of explained variation based on the survival times. To allow for censored observations, they use the expected square deviation of the survival time from its expectation.

Alternatively, Kent and O'Quigley (1988) describe a measure of explained variation based on the Kullback-Leibler measure of distance or information gained. Their method is derived using a Weibull regression model that is a fully parametric proportional hazards model. They then show how the measure may be computed for a Cox PH model, which they denoted as $\tilde{\rho}_W$. An S-plus macro (*Koq*) is available from *Statlib*.

O'Quigley and Flandre (1994) proposed a measure that is based on the sum of the squared scores or Schoenfeld (1982) residuals in (9.77) for each subject under the full versus the null model. This measure can be applied to models with time-dependent covariates, but like the Schemper and Kent-O'Quigley measures, requires additional computations.

Kent and O'Quigley (1988) also suggested a simple measure of explained variation analogous to Helland's (1987) ρ_{ε^2} presented in (A.194). Given a vector of estimated coefficients $\hat{\beta}$ from the PH model with covariate vector \mathbf{X} , they suggested that an approximate measure of explained variation is

$$R_{\varepsilon^2}^2 = \frac{\hat{\beta}' \hat{\Sigma}_{\mathbf{x}} \hat{\beta}}{\hat{\beta}' \hat{\Sigma}_{\mathbf{x}} \hat{\beta} + \sigma_{\varepsilon}^2}, \quad (9.108)$$

where $\hat{\Sigma}_{\mathbf{x}}$ is the empirical estimate of the covariance matrix of the covariate vector \mathbf{X} . When the survival times are distributed as Weibull, the covariate effects can be derived from an accelerated failure-time model in $\log(T)$, where the errors are distributed as a Gumbel distribution (see Problem 9.3). In this case, $\sigma_{\varepsilon}^2 = 1.645$. However, because the PH model is distribution-free, Kent and O'Quigley suggested

using $\sigma_\varepsilon^2 = 1$ in (9.108), yielding their $R_{W,A}^2$. They showed that the latter provides an adequate approximation to the more precise measure based on the proportion of information gained, $\tilde{\rho}_W$. In the two examples presented the measure based on the estimated proportion information gain ($\tilde{\rho}_W^2$) equaled 0.56 and 0.13, whereas the approximation $R_{W,A}^2$ equaled 0.59 and 0.13, respectively.

In a widely used computer program (superseded by PROC PHREG), Harrell (1986) suggested that the proportion of explained log partial likelihood be used as a measure of explained variation, analogous to the estimated entropy R^2 in logistic regression, but not based on an entropy loss function. However, Schemper (1990) and Kent and O'Quigley (1988) have shown that this measure grossly underestimates the proportion of explained variation estimated from their respective measures.

Schemper (1992) conducted a simulation to assess the accuracy of simple approximations to his measure and concluded that Madalla's likelihood ratio R_{LR}^2 in (A.210) provided a satisfactory approximation to his V_2 , where the N is the total sample size, not the number of events. Over a range of settings, the median difference $V_2 - R_{LR}^2$ ranged from -0.033 to 0.003 , indicating a slight negative bias. R_{LR}^2 also provides a rough approximation to the measures of Kent and O'Quigley (1988). In the two examples presented, their measure ($\tilde{\rho}_W^2$) equaled 0.56 and 0.13, respectively, whereas the approximation R_{LR}^2 yields values of 0.49 and 0.042, respectively.

Schemper and Stare (1996) presented an extensive assessment by simulation of the properties of the Schemper, Korn-Simon, and Kent-O'Quigley measures. All but the latter were highly sensitive to the degree of censoring, whereas $\tilde{\rho}_W^2$ and its approximation $R_{W,A}^2$ were largely unaffected.

Some of these measures may also be used to assess the partial R^2 or partial variation explained by individual covariates or sets of covariates, adjusted for other factors in the model. The simple approximation R_{LR}^2 would use the likelihood ratio chi-square statistic for the contribution of the covariate(s) to the full model. Kent and O'Quigley's measure of the proportion of variation explained by a covariate, say the first $j = 1$, is computed as

$$R_{W,A(1)}^2 = \frac{\hat{\beta}'_1 \hat{\Sigma}_{11.2} \hat{\beta}_1}{\hat{\beta}'_1 \hat{\Sigma}_{11.2} \hat{\beta}_1 + 1}, \quad (9.109)$$

where $\hat{\Sigma}_{11.2} = \hat{\Sigma}_{11} - \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1} \hat{\Sigma}_{21}$ is the conditional variance $\hat{V}(x_1 | \mathbf{x}_2)$ for $\mathbf{X}_2 = (X_2, \dots, X_p)$. Schemper's V_2 , however, is a measure of the variation in the survival probabilities that is explained by the full model. It does not describe the contribution of the individual covariates.

All of the above measures apply to models with baseline (fixed) covariates. For models with time-dependent covariates, Schemper's V_2 and the approximation R_{LR}^2 may be applied. However, it is not clear whether the Kent-O'Quigley approximation $R_{W,A}^2$ would apply in this case.

For a stratified PH model, there is no suitable measure of the proportion of variation explained by the stratification levels because the background hazards for each strata do not enter into the model directly.

9.4.10 SAS PROC PHREG

SAS PROC PHREG provides many of the foregoing computations. SAS version 9.2 (2008a), the latest available at the time of this writing, includes many enhancements relative to earlier versions. The syntax of the model specification is similar to that of LIFETEST using a statement of the form

```
model time*censor(0)=x1 x2 / covb corrb risklimits;
```

to indicate that the hazard function is modified proportionately by the effects of the covariates x_1 and x_2 . The program will provide likelihood ratio and score tests of the model and Wald tests of the coefficients. The *covb* and *corrb* options print the model-based estimates of the covariance matrix of the coefficient estimates and their correlation. The *risklimits* option prints the estimated hazard ratio per unit increase in the value of the covariate obtained as $\exp(\hat{\beta})$ and the corresponding 95% confidence limits. The program does not provide type III likelihood ratio tests of covariate effects; however, these can be computed by hand by fitting the appropriate nested models and computing the difference in the model chi-square values.

Alternately, each observation may be assigned a start and stop time of exposure, say T_1 and T_2 , called *counting process notation* (see example 9.7), in which case the model statement would be of the form.

```
model (T1, T2)*censor(0)=...
```

The *class* statement allows direct specification of effects, such as x_1*x_2 to designate the interaction of two variables. Unlike PROC LOGISTIC, however, the default is reference (*ref*) coding with the highest-numbered (or alphanumeric) category as the reference. However, this can be altered using the *ref* option. For example, if x_1 is a binary (0, 1) variable, where 0 is the desired reference, then a *model* statement with the *class* statement

```
class x1 (ref="0");
```

will generate the log hazard ratio for categories 1:0. Other options are *ref=first* or *ref=last* ordered value as the reference.

To allow for ties, the program provides four options of the form *ties=exact* or *Breslow* or *Efron* or *Discrete*. The *exact* option fits the exact marginal likelihood (see Kalbfleisch and Prentice, 1973, 1980), whereas the *Breslow* (the default) option fits the Peto-Breslow approximation and the *Efron* option fits an approximation from Efron (1977). The *discrete* option fits the Cox discrete logistic model, which is more appropriate when the times are coarsely grouped or are discrete. The others are more appropriate when the times may be tied because of rounding of the event times, such as to the day or week.

The *strata* option fits a stratified model as described in Section 9.4.2. Programming statements can also be used to define covariate values, depending on the values of the

strata effects or on time. Thus, it accommodates time-dependent covariates of any form.

Later versions of the program that allow a counting process data structure also provide the multiplicative intensity model for recurrent events. The program also computes the Lin and Wei (1989) robust information sandwich and the robust score test. Extensions are described in subsequent sections.

Many of the program features are illustrated in the following examples.

Example 9.6 *Squamous Cell Carcinoma (continued)*

The analyses in Example 9.1 were restricted to the subset of patients who were nonambulatory. The complete data set includes the age of the patient and the indicator variable for performance status (*perfstat*: 0 if ambulatory, 1 if not). Treatment is denoted by the value of *group*: 0 if *B*, 1 if *A*. The complete data set includes 194 patients, of whom 127 showed spread of disease during follow-up and 67 had right-censored event times. The data set is available from the book website (Squamous).

An overall unadjusted assessment of the treatment group effect would employ the SAS statement:

```
proc phreg; model time*delta(0) = group / risklimits;
```

that would produce the following output:

```
Testing Global Null Hypothesis: BETA=0
      Without      With
Criterion Covariates Covariates Model Chi-Square
-2 LOG L      1085.464    1084.090  1.374 with 1 DF (p=0.2411)
Score          .          .      1.319 with 1 DF (p=0.2507)
Wald          .          .      1.313 with 1 DF (p=0.2519)

Analysis of Maximum Likelihood Estimates
      Parameter Standard Wald      Pr >
Variable DF   Estimate  Error   Chi-Square Chi-Square
GROUP     1    -0.25074  0.21889   1.3122    0.2520

Hazard    95% Confidence Limits
Ratio      Lower      Upper
0.778     0.507     1.195
```

This shows that the overall hazard or risk ratio for treatment *A:B* is 0.778, which is not statistically significant with $p \leq 0.25$ by the likelihood ratio test.

In a model fit using only the two covariates (not shown), age is not statistically significant but the effect of performance status is highly significant ($p \leq 0.0045$ by a Wald test). Those who are not ambulatory have a risk 1.7 times greater than those who are ambulatory. The likelihood ratio model chi-square value is 8.350 on 2 *df*.

An additional model was fit using the following statements to provide an estimate of the treatment group effect adjusted for the effects of age and performance status:

```
proc phreg; model time*delta(0) = age perfstat group
/ risklimits;
```

This yields the following results:

Testing Global Null Hypothesis: BETA=0					
	Without	With			
Criterion	Covariates	Covariates	Model	Chi-Square	
-2 LOG L	1085.464	1073.951	11.513	with 3 DF	(p=0.0093)
Score	.	.	11.766	with 3 DF	(p=0.0082)
Wald	.	.	11.566	with 3 DF	(p=0.0090)

Analysis of Maximum Likelihood Estimates					
	Parameter	Standard	Wald	Pr >	
Variable	DF	Estimate	Error	Chi-Square	Chi-Square
AGE	1	-0.01067	0.00933	1.3077	0.2528
PERFSTAT	1	0.59537	0.19015	9.8039	0.0017
GROUP	1	-0.38733	0.22522	2.9576	0.0855

Hazard	95% Confidence Limits	
Ratio	Lower	Upper
0.989	0.971	1.008
1.814	1.249	2.633
0.679	0.437	1.056

The likelihood ratio test for the addition of the treatment group effect, computed by hand is $X^2 = 11.513 - 8.35 = 3.163$ with $p \leq 0.0753$, slightly more significant than the Wald test for treatment group.

The proportion of variation in the empirical survival functions explained by the model using Schemper's $V_2 = 0.055$. The approximate variation explained by the model is $R^2_{LR} = 1 - \exp(-11.513/194) = 0.058$. Based on the likelihood ratio test for the effect of treatment, the approximate proportion of variation explained by treatment group is $R^2_{LR} = 1 - \exp(-3.163/194) = 0.016$. The Kent and O'Quigley (1988) approximation to their measure of explained information gained by the full model is $R^2_{W,A} = 0.105$, and that explained by treatment group is 0.027. Because these measures are not affected by censoring, they are somewhat larger than the Schemper and R^2_{LR} measures that are reduced by censoring.

An additional model was fit with pairwise interactions among the three covariates in the linear exponent. The Wald test of the interaction between treatment group and performance status had a value $X^2 = 3.328$ with $p \leq 0.0681$, which suggests some heterogeneity of treatment group effects between the performance status groups. Dropping the other nonsignificant interactions, an additional model was fit with a treatment group effect nested within the levels of performance status; i.e., using variables *group0* and *group1* defined separately within each level of performance status as in the following statements:

```
proc phreg;
model time*delta(0) = age perfstat group0 group1 / risklimits;
group0 = (1-perfstat)*group; group1 = perfstat*group;
```

Among those who were ambulatory (*perfstat* = 0) the estimated relative risk for treatment group *A:B* is 0.996 with $p \leq 0.99$. However, among those who were not ambulatory (*perfstat* = 1), the estimated relative risk is 0.436 with $p \leq 0.0175$.

Since the treatment effect appears to depend on the value of performance status, a final model was fit that was stratified by performance status and which had nested effects for age and treatment group within strata using the likelihood (9.73). This model allows the background hazard function and the covariate effects to differ between performance status strata. The model is fit using the statements

```
proc phreg;
model time*delta(0) = age0 age1 group0 group1 / risklimits;
strata perfstat;
age0 = (1-perfstat)*age; age1 = perfstat*age;
group0 = (1-perfstat)*group; group1 = perfstat*group;
```

Note the use of programming statements to define the respective nested effects.

The resulting computations are

Testing Global Null Hypothesis: BETA=0					
	Without		With		
Criterion	Covariates	Covariates	Model	Chi-Square	
-2 LOG L	918.810	912.486	6.324	with 4 DF (p=0.1762)	
Score	.	.	5.861	with 4 DF (p=0.2097)	
Wald	.	.	5.751	with 4 DF (p=0.2185)	

and

Analysis of Maximum Likelihood Estimates					
	Parameter	Standard	Wald	Pr >	
Variable	DF	Estimate	Error	Chi-Square	Chi-Square
AGE0	1	-0.00650	0.01229	0.2799	0.5967
AGE1	1	-0.02195	0.01491	2.1676	0.1409
GROUP0	1	0.03779	0.29517	0.0164	0.8981
GROUP1	1	-0.77003	0.37309	4.2597	0.0390

Hazard	95% Confidence Limits	
Ratio	Lower	Upper
0.994	0.970	1.018
0.978	0.950	1.007
1.039	0.582	1.852
0.463	0.223	0.962

This analysis shows that the treatment group effect is nominally significant at the 0.05 level within the subgroup of patients who are not ambulatory with an estimated relative risk of 0.463 with $p \leq 0.039$, whereas there is no evidence of any treatment effect within the nonambulatory subgroup with an estimated relative risk of 1.04.

Unfortunately, it is not possible to conduct a direct test of the difference between the stratified model and that which is not stratified by performance status because the models are not nested in the coefficient parameters and thus the likelihoods are not comparable. Thall and Lachin (1986) describe a visual assessment of the hypothesis of homogeneous background hazards within strata.

If the protocol had specified that the primary analysis was the comparison of the event-free or cumulative incidence curves, then the principal analysis would be the logrank test supplemented by an estimate of the relative hazard (relative risk) using either the Mantel-Haenszel estimate, the score-based estimate, or the estimate obtained from the Cox proportional hazards model. The latter is fully efficient under the proportional hazards model and is a consistent estimate of the time-averaged relative hazard when the proportional hazards model does not apply. In the latter case, however, the model-based confidence limits are too narrow, and thus limits based on the robust information sandwich estimate of the covariance matrix are preferred.

For this study, the logrank test applied to the complete sample of 194 patients yields a Mantel-logrank test value of $X^2 = 1.372$ with $p \leq 0.25$. The Peto score-based estimate of the relative risk is $\widehat{RR} = 0.783$ with 95% confidence limits (0.52, 1.18). These are close to the Cox model-based estimates from the first of the above models.

For this study, however, it is also important to note the suggestion of heterogeneity of the treatment effect within strata defined by the performance status of the patients. In this case, especially if the objective of the study were to conduct exploratory analyses of these relationships, it would be most appropriate to report the final model that is stratified by ambulatory status with stratum-specific covariate effects.

Example 9.7 Robust Information Sandwich

To illustrate application of the Lin-Wei robust covariance matrix estimate, consider the no-interaction model with treatment group adjusted for age and performance status. This would be provided by the following statements:

```
proc phreg data = intimes covm covb;
model (TStart,TStop)*delta(0) = age perfstat group
/ risklimits covb;
```

The *covm* and *covb* options provide the following model-based estimate of the covariance matrix of the estimates obtained as the inverse of the information matrix:

Model-Based Covariance Estimate			
	AGE	PERFSTAT	GROUP
AGE	0.0000869892	-0.0001999262	0.0001887519
PERFSTAT	-0.0001999262	0.0361551619	-0.0085783204
GROUP	0.0001887519	-0.0085783204	0.0507238140

This yields the model-based standard errors for the three effects as stated above, specifically the values 0.00933, 0.19015 and 0.22522, respectively, and the resulting Wald chi-square values and the 95% confidence limits.

The *covs* and *covb* options then provide the robust estimate of the covariance matrix:

Sandwich Covariance Estimate			
	AGE	PERFSTAT	GROUP
AGE	0.0000809379	-0.0003169427	0.0002509349
PERFSTAT	-0.0003169427	0.0342851357	0.0017140566
GROUP	0.0002509349	0.0017140566	0.0516195894

This matrix yields Wald tests and confidence limits for the model coefficients that are similar to the model-based estimates.

Analysis of Maximum Likelihood Estimates						
with Sandwich Variance Estimate						
Parameter	Standard	StdErr	Ratio	Chi-Square	ChiSq	Pr >
Parameter	DF	Estimate	Error	Ratio	Chi-Square	ChiSq
AGE	1	-0.01067	0.00900	0.965	1.4055	0.2358
PERFSTAT	1	0.59537	0.18516	0.974	10.3387	0.0013
group	1	-0.38733	0.22720	1.009	2.9063	0.0882
Hazard 95% Confidence Limits						
Ratio Confidence Limits						
0.989	0.972	1.007				
1.814	1.262	2.607				
0.679	0.435	1.060				

The sandwich covariance estimate provides standard errors close to those provided by the model-based covariance estimate which assumes that the model variance assumptions are satisfied, as shown by values of the "StdErr Ratio" close to 1.0, each being the ratio of the sandwich S.E. estimate to the model-based estimate. The similarity of the two covariance matrices indicates that the proportional hazards model variance specification appears to be appropriate for these data.

The output also includes the global tests computed using the model-based and the sandwich covariance matrix estimates:

Testing Global Null Hypothesis: BETA=0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	11.5134	3	0.0093	
Score (Model-Based)	11.7664	3	0.0082	
Score (Sandwich)	13.1554	3	0.0043	
Wald (Model-Based)	11.5664	3	0.0090	
Wald (Sandwich)	13.8321	3	0.0031	

The score and Wald tests using the information sandwich estimate are comparable to the likelihood ratio test and score and Wald tests based on the assumed model.

Example 9.8 *Testing the Proportional Hazards Assumption*

Lin's (1991) procedure was applied to the model described in Example 9.7 using an estimating equation derived from weighted scores as in (9.107), with the Kaplan-Meier estimated survival function from the combined sample as the weights. The coefficient estimates and their estimated standard errors derived from the weighted score equation, the difference of the weighted estimates from the Cox model estimates, and the standard error of the difference were

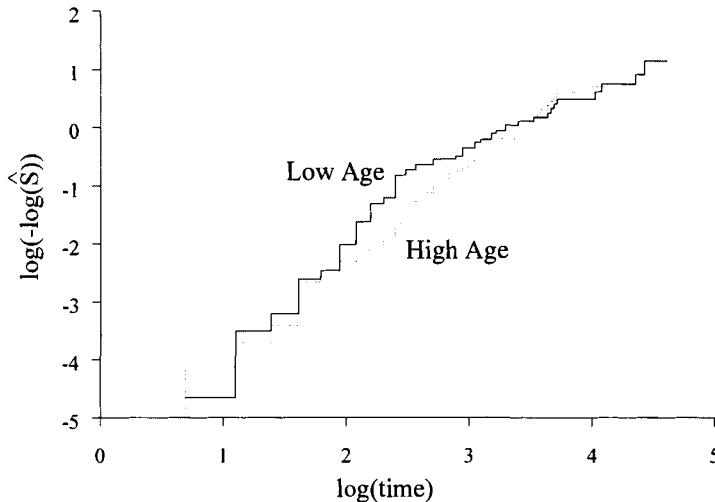
Parameter	Weighted		Difference	
	Estimate	Std. Error	Estimate	Std. Error
GROUP	-0.48504	0.251705	-0.09742	0.112351
AGE	-0.01812	0.010092	-0.00746	0.003855
PERFSTAT	0.70444	0.203207	0.10918	0.071672

The difference in the estimates of the coefficient for age is nearly twice its standard error ($Z = -0.00746/0.003855 = -1.935$), suggesting that the proportional hazards assumption for age may not apply. Using the vector of differences and its associated estimated covariance matrix (not shown) as the basis for an overall test of the three covariate proportional hazards model versus a nonproportional hazards model yields a Wald test $X^2 = 6.36$ with 3 df ($p \leq 0.0954$), which is not statistically significant.

An additional analysis can be conducted using Cox's method, in which a separate model is fit containing an interaction between each covariate, in turn, with $\log(t)$. The Wald test of the interaction effect for each coefficient did not approach significance, $p \leq 0.27$ for age, 0.15 for performance status, and 0.41 for treatment group. Cox's test, however, is designed to detect a monotone shift in the hazard ratio over time, not a general alternative.

Since Lin's analysis suggests some departure from proportional hazards for the effect of age, age was divided into two strata and a model fit for the other two covariates stratified by age. Figure 9.3 presents the plots of the $\log(-\log(\hat{S}_0(t)))$ within age strata. Noting that the failure probability is directly proportional to the complementary log-log function of the survival probability, this plot suggests (weakly) that those in the lower age stratum tend to have higher risk during the middle of the study, but equivalent risk otherwise. This nonmonotonic departure from the proportional hazards assumption was not detected by Cox's interaction test

Fig. 9.3 The Cox proportional hazards model estimated background $\log(-\log)$ survival probabilities, or the log cumulative hazard function, within the low (L) and upper (H) halves of the distribution of age, including treatment group and performance status in the model.



but was suggested by Lin's test. Overall, however, there does not appear to be a major departure from the proportional hazards assumption for this or the other covariates.

Example 9.9 Covariate-adjusted Survival Estimates

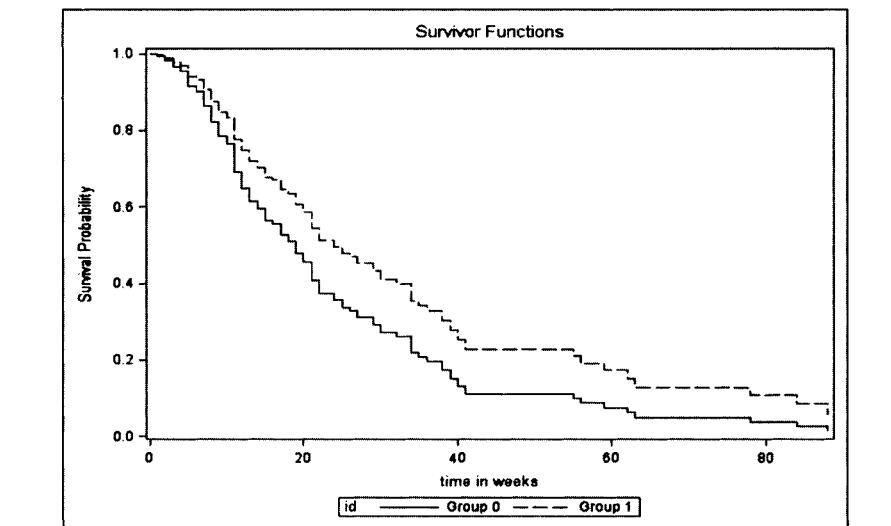
Section 9.4.7 describes the estimation of the background survival function and the computation of covariate-adjusted survival function estimates. For the squamous cell carcinoma data, the survival function estimates within each group adjusted for age and performance status would be obtained using the ODS graphics facility in PROC PHREG as follows. First, a data set *inbase* would be constructed containing one record for each group and with the variables *age* and *perfstat* set equal to their means in the cohort, such as

Obs	age	perfstat	group	id
1	59.0979	0.43814	0	Group 0
2	59.0979	0.43814	1	Group 1

The variable *id* is a character string used to identify the groups in the plot.

Then the following statements would compute the Breslow (default) estimate of the background survival function and the adjusted estimates within each group:

Fig. 9.4 Survival function within each treatment group (1 = experimental, 0 control) in the squamous cell carcinoma study from a Cox proportional hazards model using the Breslow (1972, 1974) estimated baseline survival function and adjusted for age and performance status.



```
ods graphics on;
proc phreg data = intimes plots(overlay)=survival;
model time*delta(0) = age perfstat group;
baseline covariates = inbase out = survout survival=_all_
 / rowid = id; run; ods graphics off;
```

The *plots* option specifies that a plot be generated of the survival function estimate. The *baseline* statement specifies that the function be computed using the covariate values provided within each row of the data set *inbase* and labeled by the character string in the variable *id*. The *out* = option specifies that all of the variables associated with the survival function estimates be output to the data set *survout*. This includes confidence limits that could be included as additional lines in a plot. The resulting plots are shown in Figure 9.4.

Example 9.10 DCCT Time-Dependent HbA_{1c} and Nephropathy

Example 9.2 describes the cumulative incidence of developing microalbuminuria among subjects in the Secondary Intervention Cohort of the DCCT. Such analyses established the effectiveness of intensive therapy to reduce the risk of progression of diabetic eye, kidney, and nerve disease by achieving near-normal levels of blood glucose. However, it was also important to describe the relationship between the risk of progression and the level of blood glucose control achieved, as measured by the percent of total hemoglobin that was glycosylated, or the % HbA_{1c}. The median of the current (updated) average HbA_{1c} over the period of follow-up was 8.9% among

patients in the conventional treatment group (upper and lower quartiles of 9.9% and 8%), compared to a median of 7.0% (extreme quartiles of 6.5% and 7.6%) in the intensive group. The questions were the extent to which the level of HbA_{1c} determined the risk of progression of complications, and whether the risk gradients were different between the treatment groups. These questions were addressed by the DCCT Research Group (DCCT, 1995c).

To address these questions, the DCCT Research Group (DCCT, 1995c) described Cox PH model analyses of the effects of the log of the current mean HbA_{1c}, as a time-dependent covariate, on the relative risk of developing nephropathy (microalbuminuria) in the DCCT. To fit the model, for each subject a vector (array) of updated mean HbA_{1c} values was computed as of the time of each annual visit, the values *MHBA1* to *MHBA9* for the nine years of study. Then the following SAS statements were used to fit the model separately within each treatment group:

```
proc phreg; by group;
  model time*flag(0)= lmhba / ties=discrete alpha=0.05 rl;
  array mhba (9) mhba1-mhba9;
  do j=1 to 9;
    if time eq j then lmhba=log(mhba(j));
  end;
```

Note that the observation for each subject includes the array of updated mean HbA_{1c} values at each annual visit, and additional programming statements are used to define the appropriate covariate value for each subject at each event (annual visit) time. Also, note that the array must include the values of the covariate at all event (visit) times up to the last visit for each subject. Thus, a subject with the event at year 5, or one with the last evaluation at year 5, would have the vector of values defined for years 1-5 and missing for years 6-9.

The following results were obtained from the analysis of the 316 patients in the conventional treatment group of the Secondary Intervention cohort.

Criterion	Testing Global Null Hypothesis: BETA=0		Model Chi-Square
	Without Covariates	With Covariates	
-2 LOG L	715.917	698.106	17.811 with 1 DF (p=0.0001)
Score	.	.	17.646 with 1 DF (p=0.0001)
Wald	.	.	17.364 with 1 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates					
Variable	DF	Parameter	Standard	Wald	Pr >
		Estimate	Error	Chi-Square	Chi-Square
LMHBA	1	3.170502	0.76086	17.36408	0.0001

Because the log HbA_{1c} was used in the model, then $e^{\hat{\beta}} = e^{3.17} = 23.8$ is the risk ratio per unit change in the log(HbA_{1c}), that is not very meaningful. Rather, it is more informative to consider the change in risk associated with a proportionate

change in HbA_{1c} . Since $\beta = \log(RR)$ per $\Delta \log X(t) = 1$, it can be shown (see Problem 9.14.11) that $100(c^\beta - 1)$ represents the *percentage* change in risk per a c -fold change in $X(t) = \text{HbA}_{1c}(t)$, ($c > 0$). Therefore, the estimated coefficient $\hat{\beta} = 3.17$ represents a 35.3% increase in the risk of developing nephropathy per 10% higher value of the current mean HbA_{1c} at any point in time ($c = 1.1$), or a 28.4% decrease in risk per a 10% lower HbA_{1c} ($c = 0.9$). Using the 95% confidence limits for β yields 95% confidence limits for the risk reduction per a 10% lower HbA_{1c} of (16.2, 38.8%).

Among the measures of explained variation described in Section 9.4.9, only the crude approximate measure R_{LR}^2 may be readily applied because the model included time-dependent covariates. Based on the likelihood ratio chi-square test, the log of the current mean HbA_{1c} explains $100[1 - \exp(-17.811/316)] = 5.48\%$ of the variation in risk. In DCCT (1995c), such measures were used to describe the relative importance of different covariates, not as an absolute measure of explained variation.

In analyses of the total conventional group, stratified by primary and secondary cohort, and adjusting for the baseline level of the log albumin excretion rate, the estimated coefficient $\hat{\beta} = 2.834$ corresponds to 25.8% risk reduction per 10% lower HbA_{1c} , with similar risk reductions in the primary and secondary intervention cohorts (DCCT, 1995c). Nearly equivalent results were obtained in the intensive treatment group, $\hat{\beta} = 2.639$. Thus, virtually all of the difference between the treatment groups in the risk of developing microalbuminuria was attributable to the differences in the level of glycemia as represented by the HbA_{1c} . An additional paper (DCCT, 1996) showed that there is no threshold of glycemia (HbA_{1c}) below which there is no further reduction in risk.

9.5 EVALUATION OF SAMPLE SIZE AND POWER

9.5.1 Exponential Survival

In general, a distribution-free test such as the Mantel-logrank test is used for the analysis of survival (event-time) data from two groups. In principle, the power of such a test can be assessed against any particular alternative hypothesis with hazard functions that differ in some way over time. It is substantially simpler, however, to consider the power of such tests under a simple parametric model. The Mantel-logrank test is the most commonly used test in this setting, which is asymptotically fully efficient against a proportional hazards or Lehmann alternative. The simplest parametric form of this model is the exponential model with constant hazard rates λ_1 and λ_2 over time in each group.

Asymptotically, the sample estimate of the log hazard rate $\log(\hat{\lambda})$ is distributed as $\mathcal{N}[\log(\lambda), E(d_\bullet|\lambda)^{-1}]$. Thus, the power of the test depends on the expected total number of events $E(d_\bullet|\lambda)$ to be observed during the study. Here $E(d_\bullet|\lambda) = NE(\delta|\lambda)$, where δ is a binary variable representing observation of the event ($\delta = 1$) versus right censoring of the event time ($\delta = 0$); and $E(\delta|\lambda)$ is the probability that the event will be observed as a function of λ and the total exposure of the cohort

(patient years of follow-up). The test statistic then is $T = \log(\hat{\lambda}_1/\hat{\lambda}_2)$. Under H_0 : $\lambda_1 = \lambda_2 = \lambda$, the statistic has expectation $\mu_0 = \log(\lambda_1/\lambda_2) = 0$, while under H_1 , $\mu_1 = \log(\lambda_1) - \log(\lambda_2)$. As in Section 3.3.1, let ξ_i refer to the expected sample fraction in the i th group ($i = 1, 2$) where $E(n_i) = N\xi_i$. Then the variance of the test statistic under the alternative hypothesis is

$$V(T|H_1) = \sigma_1^2 = \frac{1}{N} \left[\frac{1}{\xi_1 E(\delta|\lambda_1)} + \frac{1}{\xi_2 E(\delta|\lambda_2)} \right], \quad (9.110)$$

and under the null hypothesis is

$$V(T|H_0) = \sigma_0^2 = \frac{1}{N} \left[\frac{1}{E(\delta|\lambda)} \left(\frac{1}{\xi_1 \xi_2} \right) \right]. \quad (9.111)$$

The resulting basic equation relating sample size N and power $Z_{1-\beta}$ is

$$\begin{aligned} \sqrt{N} |\log(\lambda_1) - \log(\lambda_2)| &= Z_{1-\alpha} \left[\frac{1}{E(\delta|\lambda) \xi_1 \xi_2} \right]^{1/2} \\ &+ Z_{1-\beta} \left[\frac{1}{\xi_1 E(\delta|\lambda_1)} + \frac{1}{\xi_2 E(\delta|\lambda_2)} \right]^{1/2}, \end{aligned} \quad (9.112)$$

where $\lambda = \xi_1 \lambda_1 + \xi_2 \lambda_2$ analogously to (3.35) in the test for proportions. When expressed in terms of expected numbers of events, this yields

$$|\log(\lambda_1/\lambda_2)| = Z_{1-\alpha} \sqrt{\frac{1}{E(d_{\bullet i(0)})} + \frac{1}{E(d_{\bullet i(0)})}} + Z_{1-\beta} \sqrt{\frac{1}{E(d_{\bullet i(1)})} + \frac{1}{E(d_{\bullet i(1)})}}, \quad (9.113)$$

where $E(d_{\bullet i(0)})$ is the expected number of events in the i th group under H_0 as a function of the common assumed value λ , and where $E(d_{\bullet i(1)})$ is the like value under H_1 as a function of the assumed λ_i . These in turn are a function of the corresponding probabilities $E(\delta|\lambda)$, $E(\delta|\lambda_1)$, and $E(\delta|\lambda_2)$ that are a function of the study sample size and period of follow-up, among other factors.

Lachin (1981) and Lachin and Foulkes (1986) present a similar expression for the case where the test statistic is the difference in estimated hazard rates, $T = \hat{\lambda}_1 - \hat{\lambda}_2$. Freedman (1982) shows that these expressions can also be derived from the null and alternative distributions of the Mantel-logrank statistic. As cited by Lachin and Foulkes (1986), computations of sample size and power using the difference in hazards are conservative relative to those using the log hazard ratio in that the former yields larger required sample sizes and lower computed power for the same values of λ_1 and λ_2 . As the difference in hazards approaches zero, the difference in the two methods also approaches zero. In some respects, use of the difference in hazards would be preferred because, in fact, the Mantel-Haenszel (logrank) test statistic can be expressed as the weighted sum of the difference in the estimated hazards between groups (see Section 9.7.3). Herein, however, we use the log hazard ratio because generalizations also apply to the Cox PH model.

In the simplest case of a study with no censoring of event times, $E(\delta|\lambda) = 1$ and the event times of all subjects are observed ($N = d_*$). In this case, the total number of events d_* and power $Z_{1-\beta}$ are obtained as the solutions to

$$\sqrt{d_*} |\log(\lambda_1) - \log(\lambda_2)| = \frac{Z_{1-\alpha} + Z_{1-\beta}}{\sqrt{\xi_1 \xi_2}}. \quad (9.114)$$

Thus, the total number of events d_* required to provide power $1 - \beta$ to detect a specified hazard ratio in a test at level α is

$$d_* = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\xi_1 \xi_2 [\log(\lambda_1/\lambda_2)]^2}, \quad (9.115)$$

(George and Desu, 1974; Schoenfeld, 1981). Usually, however, there are censored event times because of administrative curtailment of follow-up (administrative censoring) or because of random losses to follow-up.

To allow for administrative censoring, let T_S designate the maximum total length of study. In the simplest case, each subject is followed for T_S years of exposure and $E(\delta|\lambda) = 1 - e^{-\lambda T_S}$. Typically, however, patients enter a study during a period of recruitment of T_R years and are then followed for a maximum total duration of T_S years ($T_S \geq T_R$) so that the first patient entered is followed for T_S years and the last patient for $T_S - T_R$ years. In a study with uniform entry over the recruitment interval of T_R years and with no random losses to follow-up, it is readily shown (see Problem 9.1.5) that

$$E(\delta|\lambda) = 1 - \frac{e^{-\lambda(T_S - T_R)} - e^{-\lambda T_S}}{\lambda T_R}, \quad (9.116)$$

(Lachin, 1981). Substitution into (9.112) and solving for N yields the sample size needed to provide power $1 - \beta$ in a test at level α to detect a given hazard ratio, say $N(\lambda)$.

Rubenstein et al. (1981) and Lachin and Foulkes (1986) present generalizations that allow for randomly censored observations because of loss to follow-up. Let γ_i be an indicator variable that denotes loss to follow-up at random during the study prior to the planned end at T_S . If we assume that times of loss are exponentially distributed with constant hazard rate η over time, then for a study with uniform entry the probability of the event is

$$E(\delta|\lambda, \eta) = \frac{\lambda}{\lambda + \eta} \left[1 - \frac{e^{-(\lambda + \eta)(T_S - T_R)} - e^{-(\lambda + \eta)T_S}}{(\lambda + \eta)T_R} \right], \quad (9.117)$$

and the probability of loss to follow-up is

$$E(\gamma|\lambda, \eta) = \frac{\eta}{\lambda} E(\delta|\lambda, \eta) \quad (9.118)$$

(see Problem 9.1.6). When $\eta_1 = \eta_2$ for the two groups, the equation relating sample size and power is obtained by substituting $E(\delta|\lambda, \eta)$ in (9.112) evaluated at λ_1, λ_1 ,

and λ . When $\eta_1 \neq \eta_2$, the probability of the event in each group under H_0 will differ. In this case, the general equation (9.112) is modified to employ the term

$$Z_{1-\alpha} \left[\frac{1}{\xi_1 E(\delta|\lambda, \eta_1)} + \frac{1}{\xi_2 E(\delta|\lambda, \eta_2)} \right]^{1/2}. \quad (9.119)$$

In the more general case, where losses to follow-up may not be exponentially distributed, Lachin and Foulkes (1986) show that the sample size with losses that follow any distribution $G(t)$ can be obtained approximately as

$$N[\lambda, G(t)] \doteq N(\lambda) \frac{E(\delta|\lambda)}{E[\delta|\lambda, G(t)]}, \quad (9.120)$$

where $E[\delta|\lambda, G(t)]$ is the probability of an event during the study. Thus, random losses to follow-up, whatever their distribution, that result in a 10% reduction in the probability of the event being observed require that the sample size needed with no losses to follow-up, $N(\lambda)$, be increased (1/0.9)-fold, or by 11.1%.

For cases where the exponential model is known not to apply, Lakatos (1988) describes the assessment sample size based on the power function of a weighted Mantel-Haenszel test against an arbitrarily specified alternative hypothesis. In this procedure one specifies the hazard rates in the control and experimental groups over intervals of time along with other projected features of the study, such as the proportion censored or lost to follow-up within each interval. This allows the assessment of power under any alternative, including cases where the hazards may not be proportional, or may even cross, and where the pattern of random censoring is uneven over time and may differ between groups. Wallenstein and Wittes (1993) also describe the power of the Mantel-logrank test in an analysis where the hazards need not be constant nor proportional over time, although the Lakatos procedure is more general.

SAS PROC POWER provides the assessment of sample size or power using the Lakatos procedure. For constant hazards over time, the results are equivalent to those obtained from the above expressions.

9.5.2 Cox's Proportional Hazards Model

Schoenfeld (1983) also derived an expression like (9.115) from the distribution of the score equation for a single binary covariate in the Cox PH model assuming equal censoring distributions within the two treatment groups. His derivation was then expanded by Hsieh and Lavori (2000) to derive the power function of the test for a single quantitative covariate.

Let x_i denote the covariate value for the i th subject. Then from (9.77) the score equation can be expressed as

$$U(\beta) = \sum_{i=1}^N \delta_i [x_i(t_i) - \bar{x}(t_i, \beta)] = \sum_{i=1}^N [s_i - E(s_i|\beta)]. \quad (9.121)$$

Let $U(\beta_0)$ denote the expression evaluated under the null hypothesis $H_0: \beta = \beta_0 = 0$, and let $I(\beta_0)$ denote the associated information. Under the alternative hypothesis

$H_1: \beta \neq 0$, the score equation $U(\beta_0)$ can be expressed as

$$U(\beta_0) = \sum_{i=1}^N [s_i - E(s_i|H_1)] + \sum_{i=1}^N [E(s_i|H_1) - E(s_i|H_0)]. \quad (9.122)$$

The first term is simply the score function $U(\beta)$ and from basic principles, $U(\beta) \sim \mathcal{N}[0, I(\beta)]$. Applying a Taylor's expansion for β in the neighborhood of $\beta_0 = 0$ to the second term, it can then be shown that

$$\sum_{i=1}^N [E(s_i|H_1) - E(s_i|H_0)] \cong \beta I(\beta_0). \quad (9.123)$$

Combining the two results, it follows that

$$U(\beta_0)|H_1 \sim \mathcal{N}[\beta I(\beta_0), I(\beta)]. \quad (9.124)$$

Thus,

$$Z|H_1 \sim \mathcal{N}\left[\beta\sqrt{I(\beta_0)}, \frac{I(\beta)}{I(\beta_0)}\right]. \quad (9.125)$$

For a specific true value β under the alternative hypothesis, the basic equation for the power of the test yields

$$\begin{aligned} |\beta|\sqrt{I(\beta_0)} &= Z_{1-\alpha} + Z_{1-\beta}\sqrt{I(\beta)/I(\beta_0)} \\ Z_{1-\beta} &= \frac{|\beta|\sqrt{I(\beta_0)} - Z_{1-\alpha}}{\sqrt{I(\beta)/I(\beta_0)}}. \end{aligned} \quad (9.126)$$

Approximately, $I(\beta) \cong I(\beta_0)$, as under a local alternative, in which case the expression for power simplifies to

$$Z_{1-\beta} = |\beta|\sqrt{I(\beta_0)} - Z_{1-\alpha}. \quad (9.127)$$

9.5.2.1 Qualitative Covariate Consider the case of a single binary (0, 1) qualitative covariate X and assuming d_{\bullet} events at distinct event times with no tied event times. Let x_i denote the covariate value for the i th subject having an event, R_i denote the risk set of those at risk at the time of the i th event, and N_i denote the number in the risk set R_i . Then it is readily shown that the score test above is equal to the Mantel logrank test with

$$I(\beta_0) = \sum_{i=1}^{d_{\bullet}} \frac{n_{0i}n_{1i}}{N_i}, \quad (9.128)$$

where (n_{0i}, n_{1i}) denote the sample sizes in the two binary categories at the i th event time. Under the null hypothesis of equal hazard functions, and with the assumption of equal censoring distributions within the two groups, then $E(n_{\ell i}) = N_i \xi_{\ell}$ for $\ell = (0, 1)$, so that $I(\beta_0) = d_{\bullet} \xi_0 \xi_1$. Under the alternative hypothesis, the log hazard ratio for categories 1:0 then is a specified value $\beta \neq 0$. Substituting into (9.127) and solving for d_{\bullet} yields

$$d_{\bullet} = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\xi_0 \xi_1 [\beta]^2}, \quad (9.129)$$

that is analogous to (9.115) derived under an exponential model.

This implies that the other methods described above for the evaluation of sample size and study duration for the logrank test under an exponential model also apply to the evaluation of the power of a test for a binary covariate in the Cox PH model. The required number of events would first be obtained from (9.129) and then the expressions in the prior section employed to determine the sample size and study duration for a specified control group hazard rate over time and rate of losses to follow-up.

9.5.2.2 Quantitative Covariate Hsieh and Lavori (2000) also apply the above to the case of a single quantitative covariate X . The score equation for β evaluated under the null hypothesis $H_0: \beta = \beta_0 = 0$ for a data set with d_\bullet events at distinct event times is

$$U(\beta_0) = \sum_{i=1}^{d_\bullet} [x_i - \bar{x}_{i(0)}] \quad (9.130)$$

$$\bar{x}_{i(0)} = \frac{\sum_{j \in R_i} x_j}{N_i}$$

and the information function is

$$I(\beta_0) = \sum_{i=1}^{d_\bullet} \left[\frac{\sum_{j \in R_i} x_j^2}{N_i} - \left(\frac{\sum_{j \in R_i} x_j}{N_i} \right)^2 \right] \quad (9.131)$$

$$= \sum_{i=1}^{d_\bullet} \frac{\sum_{j \in R_i} (x_j - \bar{x}_{i(0)})^2}{N_i} = \sum_{i=1}^{d_\bullet} V_i.$$

Since $E(V_i) = \sigma^2$ then it follows that $I(\beta_0) = d_\bullet \sigma^2$. Substituting into (9.127) yields

$$d_\bullet = \left[\frac{Z_{1-\alpha} + Z_{1-\beta}}{\beta \sigma} \right]^2. \quad (9.132)$$

In these expressions, β is the log hazard ratio per unit change in X and thus the range of risk also depends on the range or standard deviation of X . For example, assume that $\beta = 0.4$ per unit increase in X . This implies a relative risk, say R , of $e^{0.4} = 1.49$ per unit increase in X . For a $\sigma = 1$, this in turn implies a relative risk of $1.49^6 = 11.0$ over almost the complete range of X , if symmetrically distributed. However, the same β with $\sigma = 5$ implies a relative risk of $e^{0.4 \times 6 \times 5} = 162755$ over the range of X . Thus, a given β provides a wider range of risk as the range or *S.D.* of X increases.

Thus, rather than specify the minimal effect size of interest in terms of a unit difference in X , it might be clearer to describe such computations in terms of the relative hazard per standard deviation unit, say R_σ . This then yields

$$d_\bullet = \left[\frac{Z_{1-\alpha} + Z_{1-\beta}}{\sigma \ln(R_\sigma^{(1/\sigma)})} \right]^2 = \left[\frac{Z_{1-\alpha} + Z_{1-\beta}}{\ln(R_\sigma)} \right]^2. \quad (9.133)$$

For example, to detect a relative hazard of $R_\sigma = 1.49$ would require that

$$d_\bullet = \left[\frac{1.96 + 1.645}{\ln(1.49)} \right]^2 = 81.7 \quad (9.134)$$

regardless of the actual standard deviation of X . Then the above expressions under an exponential model could be employed to determine the sample size and study duration that would provide $E(d_\bullet) = 81.7$.

9.5.2.3 Adjusted Analysis Schoenfeld (1983) also showed that the expression (9.129) applies when the model includes a set of additional covariates that are assumed to be independent of the covariate X . This also applies to the expressions for a quantitative covariate. Thus, these results would apply to a comparison of two randomly assigned treatment groups adjusted for a set of baseline covariates in a randomized clinical trial.

In cases where the index binary or quantitative covariate X is not independent of a set of adjusting covariates, the adjustment described in Section 7.3.6.3 can be employed, substituting $E(d_\bullet)$ for N . Alternatively, sample size can be evaluated for a stratified analysis adjusted for other discrete (or discretized) covariates. Under an exponential model, Lachin and Foulkes (1986) describe the relationship between sample size and power for a test of the difference in hazards for two groups that is stratified by other factors. Since the logrank test is the score test in a Cox PH model, these methods also apply, approximately, to a test of the difference between groups in a Cox PH model with other binary covariates that define a set of strata.

Schoenfeld (1983) also showed that the probability of events in each of two groups under the assumption of constant proportional hazards over strata can be obtained if one has information on the survival function over time in the control group. Let $E(\delta)$ denote the resulting probability of an event in the total sample. Then the sample size required to provide power $1 - \beta$ in a test at level α to detect a given hazard ratio (R) is provided as $N = d_\bullet/E(\delta)$, where d_\bullet is the total number of events required from (9.115) substituting R for λ_1/λ_2 . Many, such as Palta and Amini (1985), describe generalizations of this approach.

9.5.2.4 Maximum Information Design For all the above computations, the actual power of a study is a direct function of the expected number of subjects reaching the event. The resulting sample size and duration for a specified control group hazard rate and rate of losses are in fact those quantities that are expected to provide the required number of subjects with an event. This approach is termed a *maximum duration design* because the study ends when the specified duration of follow-up is reached. In this case, however, the actual observed number of events is random.

Alternatively, one may adopt an "event-driven" or *maximum information design* in which subjects are recruited and followed until the required number of subjects with an event is accrued. In this case the total number of subjects with events is fixed and the duration of follow-up is random. Lachin (2005) provides a review of such designs. For a logrank test of the difference between two groups, or the test

of a binary covariate in the Cox PH model, the expression for the power function of the test is a function of $d_{\bullet}\xi_0\xi_1$ in terms of the sample fractions of subjects within the two groups or categories. This expression results under the assumption of equal censoring distributions within the two groups (categories). Alternatively, allowing for unequal patterns of censoring between the groups (categories), David Zucker (personal communication) has shown that equivalent expressions are obtained as a function of the fraction of observed events within the two groups (categories) with fractions $\zeta_i = E(d_{i\bullet}/d_{\bullet})$ for $i = 1, 2$ in (9.115) and 0, 1 in (9.129) in lieu of the group (category) sample fractions. For the latter expression, the required total information then satisfies

$$E(d_{\bullet})\zeta_0\zeta_1 = \frac{E(d_{0\bullet})E(d_{1\bullet})}{E(d_{\bullet})} = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\beta^2}.$$

To apply this approach, the right-hand side is evaluated to obtain the total information required, say I_{tot} . Then as each event is accrued, the observed information is computed as $I_{\text{obs}} = d_{0\bullet}d_{1\bullet}/d_{\bullet}$. The study is then terminated when $I_{\text{obs}} \geq I_{\text{tot}}$. Note that $I_{\text{tot}} \leq E[d_{\bullet}]/4$, so that the required total number of events is no greater than $4I_{\text{tot}}$.

Example 9.11 *Lupus Nephritis: A Study*

Lewis et al. (1992) describe the results of a clinical trial of plasmapheresis (plasma filtration and exchange) plus standard immunosuppressive therapy versus standard therapy alone in the treatment of severe lupus nephritis. The sample size for this study was based on two previous studies in which the survival function was log-linear with constant hazard approximately $\lambda = 0.3$, yielding median survival of 2.31 years. Since lupus is a rare disease, recruitment was expected to be difficult. Thus, initial calculations determined the sample size required to provide 90% power to detect 50% risk reduction (relative hazard of 0.5) with a one-sided test at level $\alpha = 0.05$. From (9.115), the total number of events required is 71.3 (rounded up to 72) with equal-sized groups and assuming no right censoring.

The study was planned to recruit patients over a period of $T_R = 4$ years with a total study duration of $T_S = 6$ years. With no adjustments for losses to follow-up, to detect a 50% risk reduction, a total $N = 127.3$ (128) subjects would be required. Among the 64 in the control group with hazard $\lambda_2 = 0.3$, the probability of the event is $E(\delta|\lambda_2 = 0.3) = 0.6804$, which yields 43.3 expected events. In the plasmapheresis group with hazard $\lambda_1 = 0.15$, the event probability is $E(\delta|\lambda_1 = 0.15) = 0.4429$ with 28.2 expected events, for a total of 71.5 (72) expected events.

This total expected number of events under the alternative hypothesis is approximately equal to the total number of events required in a study with no censoring using (9.115). Thus, the sample size is the number to be placed at risk such as to yield the required expected number of events in the presence of censoring. The greater the average period of exposure, the larger the expected number of events and the smaller the required sample size.

For example, if all patients could be recruited over a period of only six months ($T_R = 0.5$) in a six-year study, then the numbers of patients required to detect a 50% risk reduction is 101.2, substantially less than the 128 required with a four-

year recruitment period. With a six-month recruitment period, the average duration of follow-up is 5.75 years, yielding a total of 72 expected events among the 102 patients. With a four-year recruitment period, the average duration of follow-up is four years, again yielding 72 expected events among the 128 patients.

Because lupus is a life-threatening disease, it was considered unlikely that patients would be lost to follow-up at a rate greater than 0.05 per year. Thus, an additional calculation allowed for losses using (9.112) with a loss hazard rate of $\eta_1 = \eta_2 = \eta = 0.05$ in each group in (9.117). In the control group, about 11% of the patients would be lost to follow-up [$E(\gamma|\lambda_2, \eta) = 0.105$]; and the probability of observing an event is reduced to $E(\delta|\lambda_2, \eta) = 0.628$ because of these losses to follow-up. About 13% of the patients in the experimental group, $E(\gamma|\lambda_1, \eta) = 0.135$, would be lost to follow-up; and likewise the probability of an event is reduced to $E(\delta|\lambda_1, \eta) = 0.404$. Substituting these probabilities into (9.112), the total sample size is increased to $N = 138.8$ compared to $N = 127.3$ with no losses to follow-up. The total expected number of events, however, is still approximately $d_* = 72$, as before. The combined probability of the event with no losses, $0.562 = (0.680 + 0.443)/2$, is 8.8% greater than that with losses (0.516). Thus, the simple approximation in (9.120) indicates that the sample size with losses should be approximately $N = 1.088 \times 127.3 = 138.6$.

Based on such calculations, the final sample size was determined to be $N = 125$ recruited over four years with a total duration of six years. Allowing for a loss-to-follow-up hazard rate of 0.05 per year, this provides power of 0.872 to detect a 50% reduction in the hazard rate relative to the control group hazard of 0.3 per year with a one-sided test at the 0.05 level. This sample size provides power of 0.712 to detect a 40% reduction in the hazard, and only 0.553 to detect a one-third reduction. Were it not for the fact that lupus nephritis is a very rare disease, a larger sample size would have been desirable, one that would provide perhaps a minimum of 80% power to detect a one-third risk reduction.

All of the above computations relate to the design of the study as a maximum duration trial. Alternatively, a maximum information trial design would recruit and follow subjects until the required level of information is obtained. To detect a 50% risk reduction using a one-sided 0.05-level test with 90% power would require the total information

$$I_{\text{tot}} = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\beta^2} = \frac{(1.645 + 1.282)^2}{[\log(0.5)]^2} = 17.8,$$

that corresponds to no more than 72 events.

9.6 ADDITIONAL MODELS

The above developments principally apply to a single well-defined outcome event where the event and right-censoring times are also well defined in such a way that nonparametric and semiparametric methods can be developed and applied. Here we briefly describe some additional models that apply in special situations that are commonly encountered.

9.6.1 Competing Risks

Additional considerations arise when some of the censored observations are *informative* with respect to the time of the index event of interest. The most direct instance of informative censoring is the case of a *competing risk*, where each individual is at risk of an absorbing or terminal event other than the index event of interest, such as death caused by cancer in a long-term study of cardiovascular mortality. Numerous methods have been developed for examination of cause-specific risks, including the multiple decrement lifetable (cf. Elandt-Johnson and Johnson, 1980), the cause-specific hazard function (cf. Kalbfleisch and Prentice, 2002), the cause-specific subdistribution function (cf. Gray, 1988; Gaynor, Fener, Tan et al., 1993), and the conditional probability function (Pepe, 1991; Pepe and Mori, 1993), among many. More difficult problems arise when the censoring time is predictive of the ultimate failure time, but not in a deterministic manner as for a competing risk. These methods are beyond the scope of this book. However, it is important that the impact of competing risks on the analysis of event-time data be appreciated.

Consider the case where our primary objective is to describe the risk of an index event in the presence of competing risks. For example, we may wish to describe the risk of end-stage renal disease where death from nonrenal causes such as accident or cancer is a competing risk. Let T_I refer to the time of the index event and T_C refer to the time of a competing risk event. As described by Prentice et al. (1978), treating the competing risk events as unbiased censored observations provides an estimate of the *cause-specific hazard function*

$$\lambda_I(t) = \lim_{\Delta t \downarrow 0} \frac{P(t < T_I \leq t + \Delta t)}{P[(T_I \wedge T_C) > t]}, \quad (9.135)$$

where $T_I \wedge T_C = \min(T_I, T_C)$. This is the probability of the index event at time t among those at risk at that time, meaning having "survived" both the index event and the competing risk event. This is commonly called *censoring on death* because the most common application is where the index event is nonfatal, such as hospitalization, healing, remission, and so on, but where death is a competing risk. A generalized Mantel-Haenszel test then is readily constructed to test the equality of the cause-specific hazard functions for the index event in two or more groups, or a proportional hazards regression model applied to assess covariate effects on the cause-specific hazard function (see Kalbfleisch and Prentice, 2002).

In this case, however, the associated estimate of the corresponding cause-specific survival function, such as a Kaplan-Meier estimate, has no logical or statistical interpretation. For example, consider an analysis of the time to progression of renal disease where deaths from other causes are competing risks. If one treats the competing risks (nonrenal deaths) as ordinary censored observations, that is, censors on such deaths, then the estimated renal progression "survival" function provides an estimate of the probability of remaining free of progression of renal disease *in a population where there are no deaths from other causes*. Thus, the population to which the estimate applies is nonexistent. See Peto (1984), Gaynor et al. (1993),

Pepe and Mori (1993), and Mauribini and Valsecchi (1995) for other examples and further discussion.

In such instances, a much more useful quantity is the *subdistribution function*, which is the estimate of the *cumulative incidence* of the index event in a population subject to other absorbing or terminal events. The following simple adjustment provides an estimate of this function. Consider the case of a simple proportion in a cohort of patients exposed for a common unit of time in which there are no exits for any reason other than a competing risk. Let r be the number of patients entering the cohort and d be the number of events observed in a fixed unit of time. Then let e be the number of competing risk events observed, that is, events that preclude the possibility of observing the index event. As described by Cornfield (1954), in the parlance of competing risk events we can choose to estimate either of two underlying probabilities of the index event: the *net (pure) rate* as $d/(r - e)$ or the *crude (mixed) rate* as d/r . The net or pure rate estimates the probability of the event that would be observed if the competing risk event could be eliminated in the population, and thus is purely hypothetical. Conversely, the mixed rate estimates the probability of the event, or its complement, in a population at risk of the index event and also other competing risk events, analogous to the subdistribution function.

Now consider a discrete-time lifetable. Let r_j be the number entering the j th interval (i.e., at risk at the beginning of the interval); and during the j th interval let d_j be the number of index events, e_j be the number of competing risk events, and w_j be the number of unbiased randomly censored observations. To motivate a simple adjustment for the presence of competing risk events, consider the following example with three time intervals ($j = 1, 2$, or 3) and no exits for any reason other than a competing risk. Of the 100 patients entered, $\sum_j d_j = 30$ had the event and $\sum_j e_j = 20$ exited because of a competing risk event as follows:

j	r_j	d_j	e_j
1	100	5	5
2	90	10	10
3	70	15	5
<i>Total</i>		30	20

Because no patients are censored at random, the crude (mixed) proportion who survive the index event beyond the third interval is simply $(100 - 30)/100 = 0.70$. It is reasonable, therefore, to expect that a lifetable estimator over these three intervals should provide the same estimate because no patients were censored at random. The usual product-limit estimator, however, yields the estimate $(95/100)(80/90)(55/70) = 0.6635$, which is clearly incorrect. Alternatively, if competing risk exits are *not* treated as censored, but rather are retained in the lifetable throughout, then the product-limit estimator of the proportion survivors after the last interval is $(95/100)(85/95)(70/85) = 0.70$, the desired result. For such cases, in Problem 9.6.7 it is shown that this adjustment, in general, provides the proper estimate when there are no exits other than those caused by the competing risk. This approach is also discussed by Lagakos et al. (1990).

In the more general case, the subdensity function of the index event, or the probability of the index event at time t , is the probability of survival to time t and experiencing the index event at t , or

$$f_I(t) = \lambda_I(t)S_{I,C}(t), \quad (9.136)$$

where $S_{I,C}(t) = P[(T_I \wedge T_C) > t]$ is the probability of surviving both the index and competing risk events to time t and where $\lambda_I(t)$ is the cause-specific hazard for the index event at that time. Thus, the cumulative incidence or the subdistribution function of the index event is the cumulative probability

$$F_I(t) = \Pr[T_I \leq t] = \int_0^t \lambda_I(u)S_{I,C}(u) du. \quad (9.137)$$

This function may be estimated by

$$\widehat{F}_I(t) = \sum_{j:t_{(j)} \leq t} \left(\frac{d_{Ij}}{n_j} \right) \widehat{S}_{I,C}(t_{(j)}), \quad (9.138)$$

where d_{Ij} is the number of index events at the j th event time $t_{(j)}$ among the n_j subjects still at risk of either the index or competing risk events, and $\widehat{S}_{I,C}(t)$ is the Kaplan-Meier estimate of the probability of surviving both the index and competing risk event. Gaynor et al. (1993) describe a generalization of the Greenwood estimate of the variance of this function; see also Marubini and Valsecchi (1995).

Gray (1988) then presents a test for the difference between the index event subdistribution functions for two or more groups that is a generalization of the tests described in Section 9.3. Pepe (1991) also describes a test of equality of these and related functions. Lunn and McNeil (1995) show that the Cox proportional hazards model described in Section 9.4 may be used to assess covariate effects on the cause-specific hazard functions of the index and competing risk events simultaneously.

Gray also provides the package *cmprisk* for *R* that provides for the computation of the cumulative subdistribution estimates and test of significance among groups. A review of this application is provided by Scrucca et al. (2007).

It can be shown that if the number of randomly censored observations during the interval between event times is small relative to the number at risk, then the simple adjustment of not censoring on competing risks in a simple Kaplan-Meier calculation will yield estimates of the complement of the subdistribution function, $1 - \widehat{F}_I(t)$, and standard errors that are close to those provided by the above precise computations. Lagakos et al. (1990) called this computation the *intent-to-treat lifetable*. The large sample properties of this estimate, and of test statistics comparing two or more groups, relative to the more precise computations based on (9.138) have not been explored.

Finally, it should be noted that another approach is to eliminate the problem of competing risks by using a composite outcome consisting of the time to either the index or the competing risk event. In this case, inference is based on the quantity $S_{I,C}(t)$ and its associated hazard function. For example, one might use the time to progression of renal disease, transplant, or death from any cause as a composite

outcome where the observed time for each subject is the time of the first of these events to occur. Likewise, one might use all-cause mortality as an outcome rather than cause-specific mortality. Apart from the statistical simplifications, for clinical trials in cardiology, Fleiss et al. (1990) have argued that the combined outcome of fatal or nonfatal myocardial infarction (MI) is indeed more relevant than is nonfatal MI, where fatal MI is a competing risk. Their arguments also apply to an analysis of all-cause mortality versus cause-specific mortality. This approach is only feasible, however, when the component events all represent worsening of disease. In other cases, where the index event represents improvement, but some patients may worsen and ultimately die, there is no alternative other than to evaluate the cause-specific hazard, its subdistribution function, or other relevant quantities.

9.6.2 Interval Censoring

Interval censoring is the term applied to an observation where, rather than observing an event time (continuously or grouped), we only observe the boundaries of an interval of time within which an event occurred. For an individual known to experience the event, we then know that $t_i \in (a_i, b_i]$, where the subject was known to be at risk and event free at time a_i and to have had the event at some time up to time b_i . Observations that are right censored then have the associated interval $t_i \in (a_i, \infty)$. Interval censoring typically arises when the event can only be detected by performing a procedure, such as fundus photography to determine whether retinopathy has developed or progressed. In this case, a_i is the study time (e.g., day) of the last procedure where it is known that the subject is event free and b_i is the time of the procedure first conducted to determine whether the event had occurred.

Peto (1973) and Turnbull (1976) describe methods for estimation of the underlying hazard and survival functions for such data. SAS provides the *ICE* macro that can be called to perform these computations. Finklestein (1986), among others, also proposed a nonparametric test of the difference in survival functions between groups, but often without a formal demonstration of the asymptotic properties. Sun et al. (2005) describe a generalization of the Peto-Peto logrank test to interval-censored data with an explicit variance estimate and proof of the asymptotic distribution of the test. Their test includes that of Finklestein, but with a different variance estimate. The *interval* package in R provides these and related computations.

Finklestein (1986) also describes a generalization of the proportional hazards model for such data. Alioum and Commenges (1996) describe a generalization of Turnbull's estimate and also describe an extension of the PH model to such data. These and related methods all involve a large number of nuisance parameters somewhat analogous to those in the Prentice-Gloeckler model for discrete-time data.

Younes and Lachin (1997) take an alternative approach using a semiparametric model in which the background hazard function is expressed as a spline function of a smaller number of nuisance parameters. Their model is of the form

$$g[S(t|\mathbf{x})] = g[S_0(t)] + \mathbf{x}'\boldsymbol{\beta}, \quad (9.139)$$

where $S_0(t|\mathbf{x})$ is the background survival function and $g(\cdot)$ is a monotone transformation from the unit interval to the real line. The function $g(u) = \log[-\log(u)]$ yields a proportional hazards model, whereas the function $g(u) = \log[(1-u)/u]$ yields a survival proportional odds model. Both are special cases of a family of models spanned by the Aranda-Ordaz (1981) link with a parameter c ,

$$g_c(u) = \log \frac{u^{-c} - 1}{c}, \quad (9.140)$$

that yields proportional hazards as $c \rightarrow 0$ and proportional odds for $c = 1$. Other families of link functions could also be employed. The background hazard function $\lambda_0(t)$ is then expressed as a smooth function of B -splines involving a parsimonious set of nuisance parameters. The resulting likelihood function is equivalent to that for a parametric model, described below, that allows for uncensored, right-censored or interval-censored observations.

9.6.3 Parametric Models

An alternative approach is to assume an underlying parametric model that specifies a parametric distribution for the event times as a function of a rate parameter and one or more shape or scale parameters. The simplest of these is the exponential model described previously that would specify that the background hazard function is constant over time, or $\lambda_0(t) = \lambda \forall t > 0$. Feigl and Zelen (1965) describe a multiplicative exponential regression model where this constant hazard is expressed as a function of covariates, as in the Cox PH model. Pike (1966) and Peto and Lee (1983) describe a multiplicative Weibull regression model when the hazard function can be characterized by a Weibull distribution over time. The Weibull regression model is derived in Problem 9.2 for continuous-time data.

Odell et al. (1992) describe a more general Weibull regression model for interval-censored data as well as mixtures of uncensored, left-, right-, and interval-censored data. Left-censored data occurs when a subject enters the cohort having already experienced the event. Let t_i denote an observed event time, or a left or right censoring time. For interval-censored observations, let t_{Li}, t_{Ri} be the left and right censoring times, respectively. The nature of the i th observation is designated by a set of indicator variables

$$\delta_{Ri} = 1 \text{ if right censored at time } t_i \text{ (i.e., } T > t_i\text{), 0 otherwise,} \quad (9.141)$$

$$\delta_{Li} = 1 \text{ if left censored at time } t_i \text{ (i.e., } T < t_i\text{), 0 otherwise,}$$

$$\delta_{Ii} = 1 \text{ if interval censored with } t_{Li} < T \leq t_{Ri}, 0 \text{ otherwise,}$$

$$\delta_{Ei} = 1 \text{ if an event is observed exactly at } t_i, 0 \text{ otherwise,}$$

where $\delta_{Ei} = 1 - (\delta_{Ri} + \delta_{Li} + \delta_{Ii})$. For any assumed parametric distribution for the event times with density $f(t)$ and distribution function $F(t)$, the likelihood function is

$$L = \prod_{i=1}^N \{f_i(t_i)^{\delta_{Ei}} F_i(t_{Li})^{\delta_{Li}} [1 - F_i(t_{Ri})]^{\delta_{Ri}} [F_i(t_{Ri}) - F_i(t_{Li})]^{\delta_{Ii}}\} \quad (9.142)$$

(Odell et al., 1992). For a given parametric distribution, the expressions for $f(t)$ and $F(t)$ include a rate parameter $\lambda(t)$ that can then be expressed as a function of covariates, such as $\lambda(t) = \lambda_0(t) \exp(\mathbf{x}'_i \boldsymbol{\beta})$, where $\lambda_0(t)$ is a function of other parameters.

Sparling et al. (2006) generalize this model further to allow for time-dependent covariates as well as fixed baseline covariates. They do so for a family of three parameter models obtained by including a third parameter in the expression for the hazard function such that the resulting family of distributions includes the Weibull and log-logistic distributions, among others, as special cases.

For a model with fixed (e.g., baseline) covariates only, the Weibull and a wide variety of other regression models can be fit as a special case of the accelerated failure time (AFT) model. This model assumes that the distribution of the log event (survival) times is a linear function of covariates with an error distribution that is a function of the assumed underlying parametric distribution, or

$$\log(t_i) = \tilde{\alpha} + \mathbf{x}'_i \tilde{\boldsymbol{\beta}} + \sigma \varepsilon. \quad (9.143)$$

The AFT model parameters herein are designated with a \sim to distinguish them from the parameters in a model based on the parametric regression model. For example, if it is assumed that the underlying distribution is Weibull, then the error distribution is the Gumbel extreme value distribution. Problem 9.3 derives the AFT model in general, then the AFT form of a Weibull regression model and also a log-logistic model. In both cases, the transformation from AFT coefficients (and variances) to the parametric model coefficients (and variances) is described. The parametric model coefficients and variances are those obtained by substituting the parametric density and *cdf* in (9.142).

9.6.4 Multiple Event Times

In many prospective studies, multiple event times are observed of each subject (or unit), such as times to different types of events, or times to an event in multiple organs (e.g., left eye and right eye), or times to an event among members of a cluster, among many. In these settings it is desired to describe the effects of covariates on the joint risks of the component events. In the case of competing risks, the different events are each an absorbing state so that the observation of one event, say death due to cardiovascular disease, precludes observing any other event, say death due to cancer. However, when none are absorbing events it may be of interest to model covariate effects on the multiple event times marginally. For example, in previous examples we have described the incidence and prevalence of diabetic retinopathy (retinal lesions) and nephropathy (kidney dysfunction) in the DCCT. Each was assessed repeatedly over time during up to nine years of follow-up, and neither serves as a censoring mechanism for the other. Thus, an individual subject could have an event time observed for either or both outcomes.

If we desired to determine whether there is a difference between treatment groups in any of the observed outcomes, then an appropriate test would be a MANOVA-type

omnibus test of the hypothesis of no difference for all outcomes versus a difference for one or more (See Sections 4.5 and 7.9). Using the counting process construction introduced in the next section, Wei and Lachin (1984) describe a generalization of the family of weighted Mantel-Haenszel tests of such a joint null hypothesis for the difference between two groups. Palesch and Lachin (1994) generalize these tests to the case of more than two groups.

Wei et al. (1989) then describe the joint modeling of covariate effects on the risks of component event types using simultaneous Cox PH models. Consider two events, labeled as a and b , and for each we fit a Cox PH model to obtain estimated covariate vectors $\hat{\beta}_a$ and $\hat{\beta}_b$. The covariance matrix of each is provided by the estimated inverse information. The robust information sandwich is then employed to obtain the covariance matrix of the joint vector $\hat{\beta} = (\hat{\beta}_a // \hat{\beta}_b)$. Then let $\mathbf{I}(\hat{\beta}) = \text{diag} [\mathbf{I}(\hat{\beta}_a), \mathbf{I}(\hat{\beta}_b)]$ and $\mathbf{W}_i(\hat{\beta}) = [\mathbf{W}_i(\hat{\beta}_a) // \mathbf{W}_i(\hat{\beta}_b)]$. Thus, if $\hat{\beta}_a$ and $\hat{\beta}_b$ have p_a and p_b elements, p_* total, with possibly different sets of covariates in each model, then $\mathbf{I}(\hat{\beta})$ is a $p_* \times p_*$ matrix and $\mathbf{W}_i(\hat{\beta})$ is a $p_* \times 1$ vector for the i th subject with that subject's contributions to the centered score functions in (9.85) for each of the separate models. Then substitution into (9.86) and (9.87) yields the robust information sandwich covariance matrix of the joint vector estimates, $\hat{\Sigma}_R(\hat{\beta})$. This then provides for various types of tests.

For example, Wei et al. (1989) describe a clinical trial to assess the antiretroviral capacity of placebo, a low dose of ribavirin, and a high dose in AIDS patients from blood samples at 4, 8, and 12 weeks. For each collection the number of days to detection of virus positivity was the outcome. Some collections were missed or inadequate, and for some collections the time to viral positivity was right censored. Note that in this example, the outcome event variable is the same but it is reassessed up to three times. The same methods would apply if the outcome events were distinct in nature (e.g., retinopathy and nephropathy). For each of the three event times, a Cox model was fit using two binary covariates to represent the contrasts of low dose versus placebo and high dose versus placebo with coefficients $\beta_{\ell 1}$ and $\beta_{\ell 2}$ for the ℓ th event time ($\ell = a, b, c$). The 6×6 covariance matrix of the three sets of parameter estimates is then provided by the information sandwich estimate.

The omnibus (MANOVA) test of no drug effect at either dose could be tested using a robust Wald test of $H_0: \beta_{a1} = \beta_{a2} = \beta_{b1} = \beta_{b2} = \beta_{c1} = \beta_{c2} = 0$ computed as $X_O^2 = \hat{\beta}' \hat{\Sigma}_R(\hat{\beta})^{-1} \hat{\beta}$, that is asymptotically distributed as chi-square on $p_* = 6$ df. Alternatively, the effect of the low dose alone versus placebo, or $H_0: \beta_{a1} = \beta_{b1} = \beta_{c1} = 0$, could be tested using a contrast matrix

$$\mathbf{L}' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

and the quadratic form $X^2 = \hat{\beta}' \mathbf{L} [\mathbf{L}' \hat{\Sigma}_R(\hat{\beta}) \mathbf{L}]^{-1} \mathbf{L}' \hat{\beta}$, that is asymptotically distributed as chi-square on 3 df. Similarly, a separate test could be computed for the high dose versus placebo.

For this example, since the three event types are simply replicates of the same event, it would be most relevant to compute a test that the assumed common effect parameter for either dose equals zero. For the low dose this is the hypothesis $H_0: \beta_{a1} = \beta_{b1} = \beta_{c1} = \beta_{\bullet 1} = 0$. Let $\widehat{\Sigma}_{R_{\bullet 1}}$ designate the 3×3 covariance matrix of the elements corresponding to the low dose coefficient estimates alone. Then a test of this hypothesis can be based on a weighted least squares estimate of $\beta_{\bullet 1}$ and the variance of the estimate as

$$\begin{aligned}\widehat{\beta}_{\bullet 1} &= (\mathbf{J}' \widehat{\Sigma}_{R_{\bullet 1}}^{-1} \mathbf{J})^{-1} \mathbf{J}' \widehat{\Sigma}_{R_{\bullet 1}}^{-1} \widehat{\beta}_1 \\ \widehat{V}(\widehat{\beta}_{\bullet 1}) &= (\mathbf{J}' \widehat{\Sigma}_{R_{\bullet 1}}^{-1} \mathbf{J})^{-1},\end{aligned}$$

where \mathbf{J} is the ones vector and $\widehat{\beta}_1 = (\widehat{\beta}_{a1} \ \widehat{\beta}_{b1} \ \widehat{\beta}_{c1})^T$. Then the large sample test is provided by $X^2 = \widehat{\beta}_{\bullet 1}^2 / \widehat{V}(\widehat{\beta}_{\bullet 1})$ that is asymptotically distributed as chi-square on 1 df.

Lee et al. (1992) use similar steps to describe the application of the Cox model to cluster observations with variable numbers of members per cluster. They initially treat all cluster members as being independent and fit a single Cox model to describe covariate effects on the risk of an event. They then use the robust information sandwich to provide a consistent estimate of the covariance structure allowing for the correlation among members within clusters.

These models are described in the SAS/STAT *User's Guide* (2008b), and examples are provided in the SAS program *phmult.sas* that is available from the SAS/STAT library.

9.7 ANALYSIS OF RECURRENT EVENTS

One of the major advances in event-time analysis is the application of stochastic models for such counting processes to the analysis of survival and recurrent event-time data. Excellent general references are the texts by Fleming and Harrington (1991) and by Andersen et al. (1993). This approach generalizes methods for the analysis of survival data to the analysis of recurrent event data, including tests for differences between two groups of subjects and the generalized proportional hazards regression model for a counting process. A rigorous treatment is beyond the scope of this book. In the following I provide a brief description of the extensions to the analysis of recurrent events, with only a brief introduction to martingales.

Consider the case where the exact times of possibly recurrent events are recorded for each subject over a period of exposure that may vary among subjects because of staggered subject entry or random loss to follow-up. For example, such events may include successive hospitalizations, epileptic seizures, infections, accidents, or any other transient, nonabsorbing, or nonfatal event. For each subject, the sequence of times of the event constitutes a simple point process observed continuously over time (or in continuous time). From such data it is desired to estimate the underlying risk or rate function over time and to test the difference between groups in these functions.

In the more general case, we wish to assess the effects of covariates on the aggregate risk of events over time.

9.7.1 Counting Process Formulation

Consider a cohort of N subjects, where each subject is observed over a subset of the interval $(0, \tau]$. The *counting process* for the i th subject, designated as $N_i(s)$, is the number of events experienced by that subject up to study time $s \leq \tau$, for $i = 1, \dots, N$. The aggregate counting process for the cohort is $N(s) = \sum_{i=1}^N N_i(s)$, $s > 0$.

To indicate when events occur in each subject, let $dN_i(s)$ and $dN(s)$ designate the *point processes* corresponding to the "jumps" in the counting process for the i th subject or the aggregate process, respectively. Using a Stieltjes integral,

$$N(t) = \int_0^t dN(s). \quad (9.144)$$

For a continuous-time process, where it is assumed that no subject experiences multiple events at any instant and that no two subjects experience an event at the same instant, then $dN(s) = 1$ or 0 , depending on whether or not an event occurred at time s , $0 < s \leq \tau$. However, such strict continuity is not necessary. To allow for tied event times, such as multiple recurrent events in a subject on the same day, or two or more subjects with events on the same day, the point process may be defined as

$$\Delta N(s) = N(s) - N(s^-), \quad (9.145)$$

where $N(s^-) = \lim_{\Delta s \downarrow 0} N(s - \Delta s)$ is the left-hand limit at s^- , or the value of the process at the instant before time s .

To designate individuals at risk of an event at any time s , for the i th subject let $Y_i(s)$ designate the at-risk process, where $Y_i(s) = 1$ if the subject is alive and at risk (under observation) at time s ; 0 otherwise. The total number of subjects in the cohort at risk of an event at time s is $Y(s) = \sum_{i=1}^N Y_i(s)$. The at-risk indicator $Y_i(s)$ is a generalization of the simple right censoring at-risk function employed in (9.67) with the proportional hazards model, where for the i th subject followed to time t_i , $Y_i(s) = 1$ for $s \leq t_i$, 0 for $s > t_i$. In the analysis of recurrent events, however, $Y_i(s)$ may vary over time such that $Y_i(s) = 1$ over intervals of time while a subject is under surveillance and at risk and $Y_i(s) = 0$ over intervals of time while a subject is lost to follow-up or is not at risk. In this setting $Y_i(s)$ may switch back and forth between the values $0, 1$ over time. By construction, a subject is not at risk of a recurrent event over the duration of an event. For example, in an analysis of the risk of recurrent hospitalizations, $Y_i(s) = 0$ over those intervals of time while a subject is hospitalized and is not at risk of another hospitalization.

We also assume that the risk process $Y_i(s)$ is statistically independent of the counting process $N_i(s)$ for the i th subject, and likewise for the aggregate processes in the cohort. This is a generalization of the assumption of censoring at random that was employed in the analysis of survival data.

Note that the above formulation is easily adapted to the case of calendar time rather than study time. In the case of calendar time, time $s = 0$ is simply the earliest calendar time for the start of observation among all subjects, and time $s = \tau$ is the latest calendar time for the end of observation among all subjects. The $N_i(s)$, $Y_i(s)$ and $dN_i(s)$ processes are then defined in calendar time rather than in study time.

Now let \mathcal{F}_s designate the history or the *filtration* of the aggregate processes for all N subjects up to time s $\{N_i(u), Y_i(u); i = 1, \dots, N; 0 < u \leq s\}$. Likewise, \mathcal{F}_{s-} designates the history up to the instant prior to s . Then in continuous time the conditional probability (*intensity*) of an event in the i th subject at time s is

$$\alpha_i(s) ds = E [dN_i(s)|\mathcal{F}_{s-}] = \lambda(s)Y_i(s) ds \quad i = 1, \dots, N, \quad (9.146)$$

where $\lambda(s)$ is the *underlying intensity or rate function* that applies to all the subjects in the cohort. The hazard function for a single event in continuous time is now generalized to the case of possibly recurrent events as

$$\begin{aligned} \lambda(s) &= \lim_{\Delta s \downarrow 0} \frac{\Pr [\{N_i(s + \Delta s) - N_i(s) = 1\}|\mathcal{F}_{s-}]}{\Delta s} \\ &= \lim_{\Delta s \downarrow 0} \frac{\Pr [dN_i(s) |\mathcal{F}_{s-}]}{\Delta s} \end{aligned} \quad (9.147)$$

for the i th subject. Likewise, for the i th subject, the *cumulative intensity function* of the counting process $N_i(t)$ is defined as

$$\mathcal{A}_i(t) = E [N_i(t)|\mathcal{F}_t] = \int_0^t \alpha(s) ds = \int_0^t \lambda(s)Y_i(s) ds, \quad i = 1, \dots, N. \quad (9.148)$$

The intensity $\alpha_i(t)$ and cumulative intensity $\mathcal{A}_i(t)$ are interpreted as the point and cumulative expected numbers of events per subject at time t . Note that the intensity of the point process $\alpha_i(s)$ is distinguished from the underlying intensity $\lambda(s)$ in the population, in that the former allows for the extent of exposure over time within individual subjects.

Similar results apply to the aggregate counting process in the aggregate cohort. These results are now presented using Stieltjes integrals to allow for observations in continuous or discrete time. The intensity function for the aggregate point process is then defined by

$$d\mathcal{A}(s) = E [dN(s)|\mathcal{F}_{s-}] = Y(s) d\Lambda(s), \quad (9.149)$$

where, in continuous time, $d\Lambda(s) = \lambda(s) ds$, and in discrete time $d\Lambda(s) = \Delta\Lambda(s) = \Lambda(s) - \Lambda(s^-)$. Thus, the cumulative intensity function for the aggregate counting process is

$$\mathcal{A}(t) = E [N(t)|\mathcal{F}_t] = \int_0^t d\mathcal{A}(s). \quad (9.150)$$

In continuous time, $d\mathcal{A}(s) = \alpha(s) ds = \lambda(s)Y(s) ds$. In discrete time $d\mathcal{A}(s) = Y(s) \Delta\Lambda(s)$ and the integral may be replaced by a sum.

Since $N(t)$ is a nondecreasing process in time, it is termed a *submartingale*. The cumulative intensity $\mathcal{A}(t)$ is referred to as the *compensator* of the counting process

$N(t)$. Since $\mathcal{A}(t)$ is the conditional expectation of the counting process given the past history of the process \mathcal{F}_t , then the difference $\mathcal{M}(t) = N(t) - \mathcal{A}(t)$ is a *martingale* which satisfies the essential property that

$$E[\mathcal{M}(t)|\mathcal{F}_s] = \mathcal{M}(s), \quad s < t \quad (9.151)$$

or, equivalently, that

$$E[\mathcal{M}(t) - \mathcal{M}(s)|\mathcal{F}_s] = 0, \quad s < t. \quad (9.152)$$

Likewise, the change in $\mathcal{M}(t)$ at any instant in time $d\mathcal{M}(t) = dN(t) - d\mathcal{A}(t)$ is a martingale with respect to the history of the process \mathcal{F}_t . Note that $\mathcal{M}(t)$ is of the form (Observed – Expected) and thus represents the random noise associated with the observed counting process $N(t)$ over time.

The theory of martingales then provides powerful tools that establish the large sample distribution of regular functions of martingales. The variance of the counting or point process is provided by the *variation process* and the large sample distribution is described by the martingale central limit theorem.

9.7.2 Nelson-Aalen Estimator, Kernel Smoothed Estimator

We now wish to estimate the underlying intensity $\lambda(t)$ for the cohort of N individuals followed up to some time τ . Aalen (1978) presented the following estimators of the intensity for possibly recurrent events as a generalization of the estimators of the intensity for survival data first proposed by Nelson (1972). Let $\{t_{(1)} < t_{(2)} < \dots < t_{(J)}\}$ refer to the sequence of J distinct event times in the combined cohort of size N , where $dN(s) > 0$ for $s \in [t_{(1)}, \dots, t_{(J)}]$ and = 0 otherwise. We again designate the total number of subjects in the cohort at risk at the j th event time as $Y(t)$. If we assume that the intensity is discrete with probability mass at the observed event times, then from (9.149), a simple moment estimator of the jump in the cumulative intensity at time $t_{(j)}$ is

$$d\widehat{\Lambda}(t_{(j)}) = \frac{dN(t_{(j)})}{Y(t_{(j)})}, \quad 1 \leq j \leq J, \quad (9.153)$$

where $d\widehat{\Lambda}(s) = 0$ for values $s \notin [t_{(1)}, \dots, t_{(J)}]$. The corresponding *Nelson-Aalen estimator* of the cumulative intensity then is

$$\widehat{\Lambda}(t) = \int_0^t d\widehat{\Lambda}(s) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{j: t_{(j)} \leq t} \frac{dN(t_{(j)})}{Y(t_{(j)})} \quad (9.154)$$

for $0 < s \leq \tau$. Allowing for tied event times, the variance of the estimate is

$$V[\widehat{\Lambda}(t)] = \int_0^t \frac{[Y(s) - \Delta N(s)] dN(s)}{Y(s)^3}, \quad (9.155)$$

which is estimated consistently as

$$\widehat{V}[\widehat{\Lambda}(t)] = \sum_{j:t_{(j)} \leq t} \frac{dN(t_{(j)}) [Y(t_{(j)}) - dN(t_{(j)})]}{Y(t_{(j)})^3}. \quad (9.156)$$

For the analysis of survival data, or the time to the first or a single event, in the notation of Section 9.1, $d_j = dN(t_{(j)})$, $n_j = Y(t_{(j)})$, and $p_j = dN(t_{(j)})/Y(t_{(j)})$. In this case, the cumulative intensity equals the cumulative hazard and the estimate in (9.154) equals that in (9.27). Likewise, the estimate of the variance in (9.156) equals that in (9.30).

Clearly, the estimated intensity $d\widehat{\Lambda}(t)$ is a step function that can be smoothed by standard methods. The simplest approach is to use linear interpolation between successive event times such that

$$d\widehat{\Lambda}(s) = \frac{s - t_{(j)}}{t_{(j+1)} - t_{(j)}} [d\widehat{\Lambda}(t_{(j+1)}) - d\widehat{\Lambda}(t_{(j)})] + d\widehat{\Lambda}(t_{(j)}), \quad t_{(j)} < s < t_{(j+1)}, \quad (9.157)$$

where $\widehat{\Lambda}(s)$ has successive steps at $t_{(j)}$ and $t_{(j+1)}$. This is equivalent to the simple histogram method of density estimation (cf. Bean and Tsokos, 1980).

Ramlau-Hansen (1983a, b) and others have described kernel functions that yield a smoothed estimator of the intensity $\lambda(s)$. In general, the kernel estimator of the intensity is defined as

$$\widehat{\lambda}^*(s) = \frac{1}{b} \int_0^\infty K\left(\frac{s-t}{b}\right) d\widehat{\Lambda}(t) \quad (9.158)$$

with bandwidth b , where $K(u)$ is a smoothing kernel function defined for $|u| \leq 1$, 0 otherwise, and where $\int_{-1}^1 K(u) du = 1$. Computationally, the smoothed estimator is obtained as

$$\widehat{\lambda}^*(s) = \frac{1}{b} \sum_{j=1}^J K\left(\frac{s-t_{(j)}}{b}\right) \left[\frac{dN(t_{(j)})}{Y(t_{(j)})} \right]. \quad (9.159)$$

Andersen et al. (1993) describe an adjustment for values that are within b units of the left boundary ($t = 0$) that may also be applied to values approaching the right boundary ($t = t_{(J)}$).

Ramlau-Hansen (1983a) also describes a consistent estimator of the variance of this smoothed estimator. Allowing for ties, the estimated variance, say $\widehat{\sigma}^2(s) = \widehat{V}[\widehat{\lambda}^*(s)]$, is provided as

$$\widehat{\sigma}^2(s) = \frac{1}{b^2} \sum_{j=1}^J \left[K\left(\frac{s-t_{(j)}}{b}\right) \right]^2 \left[\frac{dN(t_{(j)}) [Y(t_{(j)}) - dN(t_{(j)})]}{Y(t_{(j)})^3} \right], \quad (9.160)$$

(Andersen et al., 1993). Therefore, a pointwise large sample $1 - \alpha$ confidence band is provided by $\widehat{\lambda}^*(s) \pm Z_{1-\alpha/2} \widehat{\sigma}^2(s)$.

Basically, $\hat{\lambda}^*(s)$ is a weighted average of the intensity in the neighborhood of time s using weights defined by the kernel function. If $K(u)$ is the uniform distribution, then $\hat{\lambda}^*(s)$ is the unweighted mean of the intensity over the region $s \pm b$. Usually, however, a kernel function gives weights that decline as the distance $s - t$ increases. Ramlau-Hansen (1983b) also derives a family of optimal smoothing kernels for counting process intensities. If the true intensity can be expressed as a polynomial of degree r , Ramlau-Hansen shows how to derive the optimal smoothing kernel in the sense that it minimizes the variance of the estimator of the v th derivative of the cumulative intensity function, for specified r and v . For a polynomial of degree 1, that is, a linear function in the immediate region, the optimal smoothing function for $v = 1$, that is, for the intensity itself, is the Epanechnikov (1969) kernel

$$K(u) = 0.75(1 - u^2), \quad |u| \leq 1. \quad (9.161)$$

However, as is the general case in nonparametric density estimation, the choice of the smoothing kernel is not nearly as important as the choice of the bandwidth. Although iterative procedures have been proposed for determining the optimal bandwidth, they are computationally tedious; see Andersen et al. (1993). Often, therefore, the bandwidth is determined by trial and error, starting with a bandwidth that is too wide and then successively reducing it to the point where further reductions in the bandwidth yield an overly noisy estimate. Andersen and Rasmussen (1986) present an example of kernel-smoothed intensity function estimates.

9.7.3 Aalen-Gill Test Statistics

Now consider the case where there are two groups of subjects and we wish to test the equality of the intensities

$$H_0: \lambda_1(s) = \lambda_2(s) = \lambda(s), \quad 0 < s \leq \tau, \quad (9.162)$$

where $\lambda(s) = d\Lambda(s)/ds$ is the assumed common intensity. This hypothesis can be tested using a family of nonparametric tests that are generalizations of the weighted Mantel-Haenszel test for lifetables, that includes the logrank test and the Gehan Wilcoxon test as special cases. These are also called *Aalen-Gill tests*, based on the pioneering work of Aalen (1978) and Gill (1980). Andersen (1982) and Harrington and Fleming (1982) also present further results. The properties of these tests are established using the theory of martingales for counting processes. Lan and Lachin (1995) present a heuristic introduction to martingales and their application to rank tests for survival data using the analogy of a video game originally described by Lan and Wittes (1985).

Let $N_i(s)$ and $Y_i(s)$ denote the aggregate counting and at-risk processes summed over all n_i members of the i th group, $i = 1, 2$; and let $N_*(s) = N_1(s) + N_2(s)$ and $Y_*(s) = Y_1(s) + Y_2(s)$ refer to the aggregate processes for both groups combined, all evaluated over $0 < s \leq \tau$. Then, let $K(s)$ be a *predictable process*, meaning that it is a function of the past history \mathcal{F}_{s-} , or is \mathcal{F}_{s-} -measurable, where $K(s)$ defines the weighting process for different test statistics.

The Aalen-Gill test statistic is defined as the *stochastic integral*

$$T_{AG} = \int_0^\tau K(s) \frac{Y_1(s)Y_2(s)}{Y_\bullet(s)} \left[\frac{d\mathcal{M}_1(s)}{Y_1(s)} - \frac{d\mathcal{M}_2(s)}{Y_2(s)} \right], \quad (9.163)$$

where

$$d\mathcal{M}_i(s) = dN_i(s) - d\mathcal{A}_i(s) \quad (9.164)$$

is a martingale with the compensator $d\mathcal{A}_i(s) = Y_i(s) d\Lambda(s)$ defined in terms of the assumed underlying common intensity under H_0 . Because $K(s)$, $Y_1(s)$, and $Y_2(s)$ are predictable processes, the test statistic is the difference between two stochastic integrals each of which is a *martingale transform* and thus is also a martingale. Expanding, the terms $d\Lambda(s)$ cancel to yield

$$T_{AG} = \int_0^\tau K(s) \frac{Y_1(s)Y_2(s)}{Y_\bullet(s)} \left[\frac{dN_1(s)}{Y_1(s)} - \frac{dN_2(s)}{Y_2(s)} \right]. \quad (9.165)$$

Since the Stieltjes integral can be expressed as a sum over the observed event times at which $dN_\bullet(s) > 0$, then from (9.154),

$$T_{AG} = \int_0^\tau K(s) \frac{Y_1(s)Y_2(s)}{Y_\bullet(s)} \left[d\hat{\Lambda}_1(s) - d\hat{\Lambda}_2(s) \right], \quad (9.166)$$

which is a weighted sum of the differences between the estimated intensities of the two groups.

From (9.165) the test statistic can also be expressed as

$$\begin{aligned} T_{AG} &= \int_0^\tau \frac{K(s)}{Y_\bullet(s)} [dN_1(s)Y_2(s) - dN_2(s)Y_1(s)] \\ &= \int_0^\tau K(s) \left[dN_1(s) - \frac{Y_1(s)dN_\bullet(s)}{Y_\bullet(s)} \right]. \end{aligned} \quad (9.167)$$

From (9.154), the estimate of the common underlying intensity under H_0 is

$$d\hat{\Lambda}(s) = \begin{cases} \frac{dN_\bullet(s)}{Y_\bullet(s)}, & s \in [t_{(1)}, \dots, t_{(J)}] \\ 0 & \text{otherwise} \end{cases}, \quad (9.168)$$

which is obtained from the combined sample. Thus,

$$T_{AG} = \int_0^\tau K(s) \left[dN_1(s) - \hat{E}[dN_1(s)|H_0, \mathcal{F}_{s-}] \right], \quad (9.169)$$

where the conditional expectation is

$$\hat{E}[dN_1(s)|H_0, \mathcal{F}_{s-}] = Y_1(s) d\hat{\Lambda}(s). \quad (9.170)$$

Expressing the Stieltjes integral as a sum over the observed event times $\{t_{(j)}\}$ yields

$$T_{AG} = \sum_{j=1}^J K(t_{(j)}) \left[dN_1(t_{(j)}) - \frac{Y_1(t_{(j)})dN_\bullet(t_{(j)})}{Y_\bullet(t_{(j)})} \right]. \quad (9.171)$$

Allowing for ties (cf. Fleming and Harrington, 1991; Andersen, et al., 1993), the variance of the test statistic is consistently estimated as

$$\begin{aligned}\hat{\sigma}^2 &= \int_0^\tau \frac{K(s)^2 Y_1(s) Y_2(s)}{Y_*(s)^2} \left(\frac{Y_*(s) - \Delta N_*(s)}{Y_*(s) - 1} \right) dN_*(s) \\ &= \sum_{j=1}^J \frac{K(t_{(j)})^2 Y_1(t_{(j)}) Y_2(t_{(j)})}{Y_*(t_{(j)})^2} \left(\frac{Y_*(t_{(j)}) - \Delta N_*(t_{(j)})}{Y_*(t_{(j)}) - 1} \right) \Delta N_*(t_{(j)}),\end{aligned}\quad (9.172)$$

where $\Delta N_*(t_{(j)}) = N_*(t_{(j)}) - N_*(t_{(j)}^-)$. Then from the martingale central limit theorem, asymptotically $X_{AG}^2 = (T_{AG}^2 / \hat{\sigma}^2) \stackrel{d}{\approx} \chi^2$ on 1 df under H_0 .

In terms of the notation of a 2×2 table at each event time, then $a_j \equiv dN_1(t_{(j)})$, $n_{1j} \equiv Y_1(t_{(j)})$, $m_{1j} \equiv \Delta N_*(t_{(j)})$, $N_j \equiv Y_*(t_{(j)})$, and $w(t_{(j)}) \equiv K(t_{(j)})$. Substituting into the above, the Aalen-Gill test statistic for either the time to a single event or for recurrent events is equivalent to the weighted Mantel-Haenszel test in (9.43).

The predictable weight process $K(s)$ then determines the specific form of the test. The logrank or Mantel statistic employs $K(s) = 1$ for $0 < s \leq \tau$, and the Gehan-modified Wilcoxon statistic employs $K(s) = Y_*(s)$. For recurrent events, the Harrington-Fleming G^ρ family of tests may also be defined, but in terms of the weight processes $K(s) = \exp[-\rho \hat{\Lambda}(s)]$ using the estimated cumulative intensity rather than the estimated survival function. Again $\rho = 0$ yields a logrank test, $\rho = 1$ a modified Wilcoxon-like test.

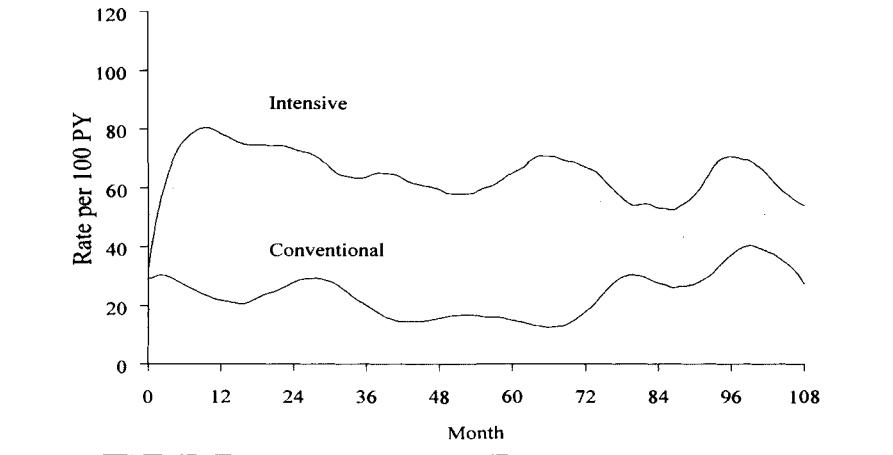
Example 9.12 DCCT Hypoglycemia Incidence

Examples in Chapter 8 describe the rate of severe hypoglycemia in the intensive and conventional treatment groups of the Diabetes Control and Complications Trial. Those results describe the overall incidence rate of possibly recurrent events over time, and the relative risk between groups, based only on the total number of events experienced by each subject in the cohort. Here we expand those analyses to incorporate information based on the time of each event. A detailed assessment of the time-varying incidence of severe hypoglycemia over time is presented by the DCCT (1997). The DCCT hypoglycemia data for the secondary intervention cohort used herein is available from the book website.

Among the 715 patients in the secondary cohort, a total of 2266 episodes of severe hypoglycemia occurred at a total of 1565 distinct event times during the total of 9.4 years of follow-up. The total number at risk ranged from 715 at the start of treatment, 713 at year 1 (365 days), 711 at year 3, 705 at year 5, 469 at year 7, and 173 at year 9. Figure 9.5 presents the kernel-smoothed estimates of the intensity (rate per 100 patient years) using a bandwidth of nine months. This shows that as intensive therapy was implemented over the first three months of treatment, the risk in the intensive group rose to a level about three times that in the conventional group. This excess risk persisted over time. The estimates beyond five years are increasingly unstable because of the declining numbers at risk.

Computation of the Aalen-Gill statistic requires that essentially a 2×2 table be constructed for each of the 1565 times at which one or more events occurred. The

Fig. 9.5 Kernel-smoothed estimates of the intensity function expressed as a rate per 100 patient years within the intensive and conventional treatment groups in the secondary cohort of the DCCT, using a bandwidth of nine months. The 95% confidence bands are also presented for each group.



test using logrank (unit) weights yields $X^2 = 562.2$ on 1 df , whereas the test using Gehan-Wilcoxon weights yields $X^2 = 564.2$, both being markedly significant.

9.7.4 Multiplicative Intensity Model

The Cox (1972) proportional hazards regression model was adapted to the problem of modeling the risks of recurrent events by Prentice et al. (1981) and Kay (1982), among others, using a partial likelihood argument. However, using the martingale theory for counting processes, Andersen and Gill (1982) describe a generalization of the PH model that can be applied to any continuous-time point process, including multiple recurrent events.

Consider the combined cohort of N subjects. For the i th subject with covariate vector $\mathbf{x}_i(s)$, possibly time dependent for $0 < s \leq \tau$, the intensity is assumed to be of the form

$$\alpha_i(s) = Y_i(s)\lambda_0(s)e^{\mathbf{x}_i(s)'\beta}; \quad 0 < s \leq \tau, \quad i = 1, \dots, N, \quad (9.173)$$

where $\lambda_0(s)$ is the background intensity of possibly recurrent events over time and $\alpha_i(s)$ is defined for those points in time when the subject is at risk of the event, $Y_i(s) = 1$. As before, let $dN_i(s)$ designate whether the i th individual experienced an event at time s ; $dN_i(s) = 1$ if yes, 0 otherwise, and denote the successive event times among all subjects as $t_{(j)}$, $1 \leq j \leq J$. The conventional partial likelihood

allowing for time-dependent covariates can then be written as

$$\tilde{L}(\beta) = \prod_{i=1}^N \prod_{j=1}^J \left[\frac{Y_i(t_{(j)}) e^{\mathbf{x}_i(t_{(j)})' \beta}}{\sum_{k=1}^N Y_k(t_{(j)}) e^{\mathbf{x}_k(t_{(j)})' \beta}} \right]^{dN_i(t_{(j)})}, \quad (9.174)$$

which is an obvious generalization of Cox's partial likelihood for survival data in (9.65), where the "risk set" of individuals at risk at each event time is explicitly defined in the denominator by the $Y_k(t_{(j)})$. The principal distinction is that subjects are removed from the risk set immediately following the event in the analysis of survival data, whereas in this model, individuals may be at risk over different periods of time and may be retained in the risk set following each event, designated by the values of $Y_i(s)$.

As for the Cox model in Section 9.4.4, the k th element of the score vector $\mathbf{U}(\beta)$ corresponding to coefficient β_k is

$$U(\beta)_{\beta_k} = \sum_{i=1}^N \sum_{j=1}^J dN_i(t_{(j)}) [x_{ik}(t_{(j)}) - \bar{x}_k(t_{(j)}, \beta)], \quad (9.175)$$

where

$$\bar{x}_k(t_{(j)}, \beta) = \frac{\sum_{\ell=1}^N Y_{\ell}(t_{(j)}) x_{\ell k}(t_{(j)}) e^{\mathbf{x}_{\ell}(t_{(j)})' \beta}}{\sum_{\ell=1}^N Y_{\ell}(t_{(j)}) e^{\mathbf{x}_{\ell}(t_{(j)})' \beta}} \quad (9.176)$$

is again the weighted average of the covariate among all those at risk at time $t_{(j)}$. The information matrix then has elements

$$\mathbf{I}_{km}(\beta) = \sum_{i=1}^N \sum_{j=1}^J dN_i(t_{(j)}) [C_{ikm}(t_{(j)}, \beta) - \bar{x}_k(t_{(j)}, \beta) \bar{x}_m(t_{(j)}, \beta)], \quad (9.177)$$

with

$$C_{ikm}(t_{(j)}, \beta) = \frac{\sum_{\ell=1}^N Y_{\ell}(t_{(j)}) x_{\ell k}(t_{(j)}) x_{\ell m}(t_{(j)}) e^{\mathbf{x}_{\ell}(t_{(j)})' \beta}}{\sum_{\ell=1}^N Y_{\ell}(t_{(j)}) e^{\mathbf{x}_{\ell}(t_{(j)})' \beta}} \quad (9.178)$$

for $1 \leq k \leq m \leq p$.

As a problem it is shown that the score test for the coefficient of a binary covariate for each of two treatment groups reduces to the Aalen-Gill generalized Mantel or logrank test.

Andersen and Gill (1982) show that the usual properties of large sample estimators and score tests are then provided by the theory of martingales for counting processes. Gill (1984) provides a readable account of these derivations with an introduction to the martingale central limit theorem. The basic result is the demonstration that the total score vector $\mathbf{U}(\beta)$ is a martingale. Even though the score vectors for the individual subjects are not *i.i.d.* with mean zero, then from the martingale central limit theorem it can be shown that asymptotically

$$\hat{\beta} \stackrel{d}{\approx} \mathcal{N}[\beta, \mathbf{I}(\beta)^{-1}], \quad (9.179)$$

where $\mathbf{I}(\boldsymbol{\beta})$ is the expected information matrix. Large sample inferences can then be obtained as for the Cox model.

There has been no exploration of measures of explained variation for use with the multiplicative intensity model for recurrent events. For lack of an alternative, the proportion of explained log likelihood, analogous to the entropy R^2 in logistic regression, may be used to assess the relative importance of alternative models or of different covariates. This measure likely underestimates the proportion of explained variation with respect to other possible metrics.

Example 9.13 DCCT Hypoglycemia Risk

Analyses were also conducted to assess the relationship between the level of HbA_{1c} as a time-dependent covariate and the risk of severe hypoglycemia over time, including the risk of recurrent events (DCCT, 1997). In both the intensive and conventional treatment groups, the dominant predictor of future hypoglycemic episodes was a history of past hypoglycemia as represented by an additional time-dependent covariate giving the number of prior episodes since entry into the trial. The analyses employed the multiplicative intensity model using the SAS PROC PHREG with "counting process" input. For example, the data for a single subject are

(start stop]	Event	$\log(\text{HbA}_{1c})$	# Prior
0 34	0	2.399712	0
34 66	0	2.198335	0
⋮	⋮	⋮	⋮
145 162	1	1.947338	0
162 189	1	1.947338	1
189 194	0	1.947338	2
⋮	⋮	⋮	⋮
600 608	1	1.891605	2
608 630	0	1.891605	3
⋮	⋮	⋮	⋮
3421 3451	0	1.960095	27

Each interval is defined by a start and stop time, closed on the right. A new interval is also constructed whenever an event occurs, in which case the value of the stop time equals the event time. A new interval is also constructed whenever the value of a time-dependent covariate changes, where the start time indicates when the change becomes effective.

In this example, the patient starts at day 0 with a $\log \text{HbA}_{1c}$ of 2.399712 and with no prior episodes of hypoglycemia. On day 34 a new blood sample was drawn for which the $\log \text{HbA}_{1c}$ value was 2.198335, and another was obtained on day 66 and periodically thereafter. As of day 145, a new $\log \text{HbA}_{1c}$ value of 1.947338 was obtained. The patient then experienced the first episode of severe hypoglycemia on day 162, followed by a second episode on day 189. On day 194 an updated value of HbA_{1c} was then obtained. On day 608 the patient experienced a third episode, and so

Table 9.6 Multiplicative Intensity model of recurrent hypoglycemia with a quadratic effect for log HbA_{1c}.

Summary of the Number of Event and Censored Values					
Stratum	PHASE2	Total	Event	Percent	
				Censored	Censored
1	0	21038	1015	20023	95.18
2	1	11375	705	10670	93.80
<hr/>					
Total		32413	1720	30693	94.69

Testing Global Null Hypothesis: BETA=0					
	Without	With			
Criterion	Covariates	Covariates	Model	Chi-Square	
-2 LOG L	17520.386	16546.349	974.037	with 3 DF	(p=0.0001)
Score	.	.	1628.964	with 3 DF	(p=0.0001)
Wald	.	.	1278.827	with 3 DF	(p=0.0001)

Analysis of Maximum Likelihood Estimates					
	Parameter	Standard	Wald	Pr >	
Variable	DF	Estimate	Error	Chi-Square	Chi-Square
LHBA1C	1	13.557041	3.94309	11.82104	0.0006
LHBA1C2	1	-3.943712	1.00977	15.25346	0.0001
NPRIOR	1	0.088835	0.00269	1090	0.0001

on. The last updated HbA_{1c} was obtained on day 3421 and the patient completed the study on day 3451, having experienced a total of 27 episodes of severe hypoglycemia during the study.

The number of prior episodes during the study is a second time-dependent covariate. It is incremented beginning with the interval immediately following that in which an event occurs. If a patient experienced more than one event on the same day, then the additional events were included in the data set as though they had occurred on successive days.

Among the 363 intensive group patients in the secondary cohort, the following SAS statements would fit a quadratic model in the log HbA_{1c}:

```
proc phreg;
strata phase2;
model (start,stop)*hypo(0) = lhba1c lhba1c2 nprior;
```

where lhba1c2 is the square of lhba1c, the log(HbA_{1c}). The model is stratified by the phase in which the subject entered the study because after the first year of recruitment (phase 2) the eligibility criteria were changed to exclude patients with a prior history of severe hypoglycemia.

The principal output of the program, including the model estimates of the coefficients, is presented in Table 9.6. Note that the data set contained 32,413 records based on data from the 715 patients who experienced a total of 1720 episodes of severe hypoglycemia.

Since a quadratic effect for $\log \text{HbA}_{1c}$ was included in the model, it can then be shown that the percentage reduction in risk for a fixed 10% reduction in the HbA_{1c} at an HbA_{1c} value of x is computed as $100(0.9^\eta - 1)$, where $\eta = \beta_1 + \beta_2 \log(0.9x^2)$. Thus, from the estimated coefficients, a 10% reduction from an HbA_{1c} of 10 yields a 55.5% increase in risk, whereas a 10% reduction from an HbA_{1c} of 8 yields a 29.2% increase in risk. Because the coefficient for the quadratic effect is negative, the risk reduction per 10% lower HbA_{1c} declines as the reference value of the HbA_{1c} declines.

As a crude measure of explained variation for the model, the proportion of log likelihood explained by these three covariates (see Section A.8.3) is provided by $R_\ell^2 = 974.037/17520.386 = 0.056$. In analyses of the complete cohort (DCCT, 1997), the risk relationships were stronger among patients in the conventional than in the intensive treatment groups, explaining a higher fraction of the log likelihood. Further, the risk gradients were different between the two groups, such that the difference in the level of glycemia alone was not the principal determinant of the difference in risk of severe hypoglycemia between the two groups.

The above computations are provided by the program *hypomim.sas*.

9.7.5 Robust Estimation: Proportional Rate Models

The essential assumption of the multiplicative intensity model, other than proportional intensity, is that the dependence among the recurrence times is captured by conditioning on covariates that influence the risk of recurrence, such that the recurrence times are conditionally independent. Of course, that is an untestable assumption. Thus, Lin et al. (2000), proposed that these assumptions be relaxed.

Stated more formally, the multiplicative intensity model is based on two assumptions:

- i. $E [dN(t)|\mathcal{F}_{t-}] = E [dN(t)|\mathbf{x}(t)]$.
- ii. $E [dN(t)|\mathbf{x}(t)] = \lambda_0(t) \exp [\mathbf{x}(t)'\boldsymbol{\beta}]$.

The first specifies that the risk of events at a point in time depends on the prior history of the process, and that this information is captured by conditioning on a time-dependent covariate vector process $\mathbf{x}(t)$. The second assumption specifies that the covariate values act upon the intensity of the process in the usual log-linear manner (i.e., proportionately). Note that if the covariate vector only includes fixed (e.g., baseline) covariates, the first assumption is equivalent to the Poisson process assumptions with independent increments, or that the time to the next event is independent of the times to prior events.

Lin et al. (2000) then proposed relaxing or dispensing with the first assumption and basing the model only on the second assumption, that is expressed as

$$E [dN(t)|\mathbf{x}(t)] = d\mu_{\mathbf{x}}(t) = d\mu_0(t) \exp [\mathbf{x}(t)'\boldsymbol{\beta}] . \quad (9.180)$$

This is termed the *proportional rate model*. Equivalently,

$$E[N(t)|\mathbf{x}(s), 0 < s < t] = \mu_{\mathbf{x}}(t) = \int_0^t e^{\mathbf{x}(s)'\boldsymbol{\beta}} d\mu_0(s). \quad (9.181)$$

If the model specifies that this expectation is a function only of baseline or fixed covariates, this yields the *proportional mean model*,

$$E[N(t)|\mathbf{x}] = \mu_{\mathbf{x}}(t) = e^{\mathbf{x}'\boldsymbol{\beta}} \mu_0(t). \quad (9.182)$$

The partial likelihood for the proportional rate or mean model is the same as that for the multiplicative intensity model. The covariance matrix of the estimates is then provided by the robust information sandwich estimate.

Thus, the coefficient estimates from the multiplicative intensity model remain unbiased, but the model-based estimate of the covariance matrix of the estimates may be affected if the first of the two assumptions is violated. Therefore, it is prudent to employ a robust information sandwich to estimate the covariance matrix of the estimates, that is, to employ the proportional rate model rather than the multiplicative intensity model.

Alternatively, Lin and Wei (1992), among others, describe a generalization of the accelerated failure-time model with multiple or recurrent events with a robust estimate of the covariance matrix of the coefficient estimates. Wei and Glidden (1997) provide a review of this and other methods for multiple failure-time data.

Example 9.14 DCCT Hypoglycemia Risk (continued)

For the above model, the robust information sandwich covariance matrix provides estimated standard errors for the three model coefficients of 4.023, 1.033, and 0.0028, respectively. These are close to those estimated from the multiplicative intensity model.

9.7.6 Stratified Recurrence Models

Prentice et al. (1981) consider a different approach by modeling the effects of covariates on the risk of a recurrent event by stratifying on the number of prior events. Two models are described, one in terms of the time from entry, and the other where time is reset to zero after each successive occurrence of the event. In the latter, the time elapsed from one event to the next is referred to as the time gap between events. Their models are essentially stratified Cox models, where the background hazard function and covariate effects may vary among strata (i.e., for each successive event).

Let t_k denote the total elapsed time from entry to the k th event occurrence. Then the model as a function of total time is expressed as

$$\lambda(t|\mathcal{F}_{t-}) = \lambda_{0k}(t) e^{\mathbf{x}(t)'\boldsymbol{\beta}_k} \quad \text{for } t_{k-1} < t \leq t_k, \quad (9.183)$$

where \mathcal{F}_{t-} refers to the observed history of the event process as described in Section 9.7.1, $\mathbf{x}(t)$ refers to the possibly time-dependent covariate values at time t , and $\boldsymbol{\beta}_k$

refers to the coefficient vector for the k th recurrence. The gap-time model would substitute $s = t - t_{k-1}$ for t in the model ($s > 0$). Note that a given subject with k observed recurrence times appears in the risk set of the first k strata up to the time of each event, and then in the $(k+1)$ st stratum until the time of right censoring. These models can be fit using PROC PHREG, but manipulation of the data set is required to properly construct the risk sets within each stratum. Examples are provided in the SAS/STAT *User's Guide* (2008b).

The essential assumption of these models, as with the multiplicative intensity model, is that the successive event times are conditionally independent given the history of the event process. Wei et al. (1989) point out that the models are sensitive to departures from this assumption and thus recommend that the robust information sandwich be used to estimate the covariance matrix of the parameter estimates.

9.8 PROBLEMS

9.1 Exponential Model. Assume a simple exponential survival model with constant hazard $\lambda(t) = \lambda$, $t > 0$, where $S(t) = e^{-\lambda t}$.

9.1.1. In a cohort of size N , none of whom have censored survival (event or death) times (i.e., all N event times are observed), derive the maximum likelihood estimator of λ and its asymptotic variance as

$$\hat{\lambda} = \frac{N}{\sum_{j=1}^N t_j}, \quad V(\hat{\lambda}) = \frac{\lambda^2}{N}, \quad (9.184)$$

in terms of the observed event times t_j ($j = 1, \dots, N$).

9.1.2. In a cohort of size N , now let there be d_{\bullet} event (deaths) observed and the remainder $N - d_{\bullet}$ right censored at random. Let δ_j be the binary indicator variable that denotes whether the j th subject's event time was observed ($\delta_j = 1$) or was right censored ($\delta_j = 0$), where $d_{\bullet} = \sum_j \delta_j$. Using the likelihood in (9.6), derive the maximum likelihood estimator and variance as

$$\hat{\lambda} = \frac{d_{\bullet}}{\sum_{j=1}^{d_{\bullet}} t_j + \sum_{j=d_{\bullet}+1}^N t_j^+}, \quad V(\hat{\lambda}) = \frac{\lambda^2}{E(d_{\bullet})} = \frac{\lambda^2}{NE(\delta)}, \quad (9.185)$$

where t_j ($j = 1, \dots, d_{\bullet}$) are the observed event times and t_j^+ ($j = d_{\bullet} + 1, \dots, N$) are the right-censored event times. Thus, asymptotically $\hat{\lambda} \sim \mathcal{N}[\lambda, \lambda^2/E(d_{\bullet})]$.

9.1.3. Under random censoring, use the delta method to show that the estimated log survival distribution at a given time t , $\log[\hat{S}(t)]$, is asymptotically normally distributed with expectation $-\lambda t$ and variance $(\lambda t)^2/E(d_{\bullet})$.

9.1.4. Also show that the estimated survival distribution at a given time t , $\hat{S}(t)$, is asymptotically normally distributed with expectation $e^{-\lambda t}$ and variance $e^{-2\lambda t}(\lambda t)^2/E(d_{\bullet})$.

9.1.5. Now consider the case of staggered entry into the trial where we assume that subjects enter the study uniformly over the interval 0 to T_R in calendar time, meaning

that the entry times, say r_j , are distributed as uniform over $(0, T_R]$. Also assume that all subjects are followed to time $T_S > T_R$, T_S being the total study duration. Then any subjects event free at the end of the study are administratively censored with a censored event time of $t_j^+ = T_S - r_j$. Allowing only for administrative censoring (i.e., no other censoring during the study), derive the expression for $E(\delta|\lambda)$ presented in (9.116).

9.1.6. Now assume that losses to follow-up occur randomly over time with a constant hazard rate η . Show that the probability of the event in the study is given by the expression in (9.117) and that the probability of being lost to follow-up is given by (9.118).

9.1.7. Consider the case of two populations with constant hazards λ_i and corresponding survival functions $S_i(t)$; $i = 1, 2$. The constant of proportionality of the hazards is $\theta = \lambda_1/\lambda_2$. Show that $S_1(t) = S_2(t)^\theta$.

9.1.8. Let $t_{i\alpha}$ be the time at which α is the remaining proportion of survivors in the i th population, or $S_i(t_{i\alpha}) = \alpha$, $0 < \alpha < 1$, $i = 1, 2$. Show that $t_{1\alpha}/t_{2\alpha} = 1/\theta$ for all $\alpha \in (0, 1)$. Thus, a hazard ratio of $\theta = 2$ implies that it will take half the time for subjects in population 1 to reach the α fraction of survivors than those in population 2; or $\theta = 1/3$ implies that it will take three times as long. Thus, when θ is the hazard ratio, $1/\theta$ is the factor by which the event times are either accelerated or decelerated.

9.2 Weibull Model. The Weibull distribution with rate parameter μ and shape parameter γ has a hazard function $\lambda(t) = \mu\gamma(t)^{\gamma-1}$, $\mu > 0$, $\gamma > 0$, where $\gamma = 1$ equals the exponential distribution.

9.2.1. Show that the Weibull survival distribution is expressed as

$$S(t) = \exp(-\mu t^\gamma) \quad (9.186)$$

and the death density as

$$f(t) = \mu\gamma(t)^{\gamma-1} \exp(-\mu t^\gamma). \quad (9.187)$$

9.2.2. For a random sample of N observations containing d_* event times $\{t_j\}$ and $N - d_*$ censored event times $\{t_j^+\}$, show that the likelihood function in the parameters $\boldsymbol{\theta} = (\mu \gamma)^T$ is

$$L(\boldsymbol{\theta}) = \prod_{j=1}^N [\mu\gamma(t_j)^{\gamma-1}]^{\delta_j} \exp(-\mu t_j^\gamma). \quad (9.188)$$

From the log likelihood and its derivatives, show that the maximum likelihood estimating equations for μ and γ are

$$\begin{aligned} U(\boldsymbol{\theta})_\mu &= \frac{d_*}{\mu} - \sum_j t_j^\gamma \\ U(\boldsymbol{\theta})_\gamma &= \frac{d_*}{\gamma} + \sum_j \delta_j \log(t_j) - \mu \sum_j t_j^\gamma \log(t_j). \end{aligned} \quad (9.189)$$

Given an estimate of γ , the *MLE* of the rate parameter is $\hat{\mu} = d_{\bullet} / \left(\sum_j t_j^{\gamma} \right)$. What does this suggest as to a starting value for μ and γ in the Newton-Raphson iteration?

9.2.3. Then show that the expected information matrix for μ and γ has elements

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta})_{\mu} &= \frac{d_{\bullet}}{\mu^2} & (9.190) \\ \mathbf{I}(\boldsymbol{\theta})_{\gamma} &= \frac{d_{\bullet}}{\gamma^2} + \mu \sum_j t_j^{\gamma} [\log(t_j)]^2 \\ \mathbf{I}(\boldsymbol{\theta})_{\mu, \gamma} &= \sum_j t_j^{\gamma} \log(t_j). \end{aligned}$$

9.2.4. Use the δ -method to show that the variance of the log estimated survival distribution at time t is

$$V \left(\log[\hat{S}(t)] \right) = t^{2\gamma} \sigma_{\mu}^2 + \mu^2 t^{2\gamma} [\log(t)]^2 \sigma_{\gamma}^2 + \mu t^{2\gamma} \log(t) \sigma_{\mu\gamma}, \quad (9.191)$$

where the variances (σ_{μ}^2 , σ_{γ}^2) and covariance ($\sigma_{\mu\gamma}$) of the parameter estimates are provided by the inverse information matrix.

9.2.5. Now let the rate parameter be a function of a vector of covariates \mathbf{x} of the form $\mu = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta})$. Show that the score equations for the elements $(\alpha, \boldsymbol{\beta})$ are

$$\begin{aligned} U(\boldsymbol{\theta})_{\alpha} &= \sum_j \left(\delta_j - e^{\alpha + \mathbf{x}'_j \boldsymbol{\beta}} t_j^{\gamma} \right) = d_{\bullet} - \sum_j \mu_j t_j^{\gamma}, \\ U(\boldsymbol{\theta})_{\gamma} &= \frac{d_{\bullet}}{\gamma} + \sum_j \delta_j \log(t_j) - \sum_j \mu_j t_j^{\gamma} \log(t_j), \end{aligned} \quad (9.192)$$

and

$$U(\boldsymbol{\theta})_{\beta_k} = \sum_j x_{jk} \left(\delta_j - e^{\alpha + \mathbf{x}'_j \boldsymbol{\beta}} t_j^{\gamma} \right) = \sum_j x_{jk} (\delta_j - \mu_j t_j^{\gamma})$$

for $1 \leq k \leq p$.

9.2.6. Show that the elements of the information matrix are

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta})_{\alpha} &= \sum_j \mu_j t_j^{\gamma}, & \mathbf{I}(\boldsymbol{\theta})_{\gamma} &= \frac{d_{\bullet}}{\gamma^2} + \sum_j \mu_j t_j^{\gamma} [\log(t_j)]^2, \\ \mathbf{I}(\boldsymbol{\theta})_{\alpha\gamma} &= \sum_j \mu_j t_j^{\gamma} \log(t_j), & \mathbf{I}(\boldsymbol{\theta})_{\beta_k} &= \sum_j x_{jk}^2 \mu_j t_j^{\gamma}, \\ \mathbf{I}(\boldsymbol{\theta})_{\alpha\beta_k} &= \sum_j x_{jk} \mu_j t_j^{\gamma}, & \mathbf{I}(\boldsymbol{\theta})_{\gamma\beta_k} &= \sum_j x_{jk} \mu_j t_j^{\gamma} \log(t_j), \\ \mathbf{I}(\boldsymbol{\theta})_{\beta_k\beta_m} &= \sum_j x_{jk} x_{jm} \mu_j t_j^{\gamma}, \end{aligned} \quad (9.193)$$

for $1 \leq k < m \leq p$.

9.2.7. Show that e^{β_j} equals the hazard ratio per unit increase in the j th covariate x_j . Thus, the Weibull model with this parameterization is a parametric proportional hazards model.

9.2.8. In this model show that the survival function for an individual with covariate vector \mathbf{x} is

$$S(t|\mathbf{x}) = \exp \left[-e^{(\alpha + \mathbf{x}'\boldsymbol{\beta})} t^{\gamma} \right] = \exp \left[-\exp(\gamma \log(t) + \alpha + \mathbf{x}'\boldsymbol{\beta}) \right]. \quad (9.194)$$

9.2.9. Then show that

$$\log(-\log[\widehat{S}(t|\mathbf{x})]) = \widehat{\gamma} \log(t) + \widehat{\alpha} + \mathbf{x}' \widehat{\boldsymbol{\beta}} \quad (9.195)$$

and that $V[\log(-\log[\widehat{S}(t|\mathbf{x})])] = \mathbf{H}' \boldsymbol{\Sigma} \mathbf{H}$, where

$$\mathbf{H} = [\log(t) \quad 1 \quad \mathbf{x}']^T. \quad (9.196)$$

The covariance matrix of the parameter estimates, $\boldsymbol{\Sigma}$, is provided by the inverse information and is partitioned as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\gamma^2 & \sigma_{\gamma\alpha} & \Sigma_{\gamma\beta} \\ \sigma_{\alpha\gamma} & \sigma_\alpha^2 & \Sigma_{\alpha\beta} \\ \Sigma_{\beta\gamma} & \Sigma_{\beta\alpha} & \Sigma_\beta \end{bmatrix}, \quad (9.197)$$

where Σ_β is a $p \times p$ matrix.

9.3 Accelerated Failure-Time Model. The accelerated failure-time model assumes that the time to any fraction of survivors is accelerated or decelerated by some function of a covariate vector \mathbf{X} with coefficient vector $\tilde{\boldsymbol{\beta}}$, usually of the form $\exp[-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma]$, where σ is a scale parameter. That is, for a specified survival function $S_0(t)$, the survival function for a patient with covariate vector \mathbf{x} is of the form

$$S(t|\mathbf{x}) = S_0(e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} t) = S_0(\tilde{t}_\mathbf{x}) \quad (9.198)$$

where $\tilde{t}_\mathbf{x} = e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} t$ is the transformed accelerated failure time determined by the covariate vector value \mathbf{x} .

9.3.1. Noting that $\Lambda(t|\mathbf{x}) = \Lambda_0(e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} t)$, show that the hazard function is of the form

$$\lambda(t|\mathbf{x}) = e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} \lambda_0(e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} t). \quad (9.199)$$

Hint: Use Leibniz's rule for the derivative of an integral. Thus,

$$f(t|\mathbf{x}) = e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} \lambda_0(e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} t) S_0(e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})/\sigma} t).$$

9.3.2. Now, let $y/\sigma = \log(t)$ so that $t = \exp(y/\sigma)$. Given any underlying distribution with hazard function $\lambda_0(t)$ and corresponding survival function $S_0(t)$, show that the conditional distribution of Y given \mathbf{x} is

$$f(y|\mathbf{x}) = e^{[y - (\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})]/\sigma} \lambda_0(e^{[y - (\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})]/\sigma}) S_0(e^{[y - (\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})]/\sigma}) (\frac{1}{\sigma}). \quad (9.200)$$

Then for $\varepsilon = [y - (\tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}})]/\sigma$ it follows that

$$f(\varepsilon) = e^\varepsilon \lambda_0(\varepsilon) S_0(\varepsilon). \quad (9.201)$$

Thus, we can adopt a linear model of the form $y_i = \tilde{\alpha} + \mathbf{x}' \tilde{\boldsymbol{\beta}} + \varepsilon_i \sigma$ with error distribution $f(\varepsilon)$.

9.3.3. Since $P(t_i > t) = P[\log(t_i) > \log(t)]$, show that

$$S(t|\mathbf{x}_i) = P\left[\varepsilon_i > \frac{y_i - (\tilde{\alpha} + \mathbf{x}'_i \tilde{\beta})}{\sigma}\right] \quad (9.202)$$

evaluated with respect to the distribution of the errors, that in turn can be obtained from the assumed underlying survival distribution $S_0(t)$. These developments can be used to obtain an accelerated failure-time model for a given parametric distribution $S_0(t)$.

9.3.4. Weibull Accelerated Failure-Time Model. In a standard Weibull model with $\mu = 1$, then $\lambda(t) = \gamma(t)^{\gamma-1}$. Adopting the accelerated failure-time transformation as in (9.198) with the acceleration factor $-(\tilde{\alpha} + \mathbf{x}' \tilde{\beta})$ for a subject with covariate vector \mathbf{x} , show that

$$\lambda(t|\mathbf{x}) = e^{-(\tilde{\alpha} + \mathbf{x}' \tilde{\beta})\gamma} \gamma(t)^{\gamma-1} \quad (9.203)$$

and that

$$S(t|\mathbf{x}) = \exp\left[-e^{-\gamma(\tilde{\alpha} + \mathbf{x}' \tilde{\beta})} t^\gamma\right] = \exp\left(-\exp\left[\log(t) - (\tilde{\alpha} + \mathbf{x}' \tilde{\beta})\right]^\gamma\right). \quad (9.204)$$

Substituting $\gamma = 1/\sigma$, it follows that

$$S(t|\mathbf{x}) = \exp\left(-\exp\left[\frac{\log(t) - (\tilde{\alpha} + \mathbf{x}' \tilde{\beta})}{\sigma}\right]\right). \quad (9.205)$$

This is the parameterization used by the SAS procedure LIFEREG that fits a Weibull accelerated failure-time model in which the covariates determine the expected value of Y and σ is the scale parameter.

Compared to (9.194) it follows that the Weibull proportional hazards parameters in Problem 9.2.5 can be obtained from the accelerated failure-time model as $\alpha = -\tilde{\alpha}/\sigma$, $\beta_j = -\tilde{\beta}_j/\sigma$, and $\gamma = 1/\sigma$. The SAS procedure LIFEREG provides estimates of $\tilde{\alpha}$, $\tilde{\beta}$, and σ and of the large sample covariance matrix of the estimates. From these, the estimates of α , β and γ are readily derived, and their covariance matrix obtained via the δ -method.

9.3.5. The Weibull accelerated failure-time model in $\log(t)$ may also be derived as follows: Let the rate parameter be expressed as a function of a covariate vector \mathbf{X} as $\mu = \exp[-(\tilde{\alpha} + \mathbf{x}' \tilde{\beta})\gamma]$, where $\gamma = \sigma^{-1}$. Then show that the distribution of $y = \log(t)$ can be expressed as

$$f(y|\mathbf{x}) = \frac{1}{\sigma} \exp\left[\frac{y - (\tilde{\alpha} + \mathbf{x}' \tilde{\beta})}{\sigma} - \exp\left(\frac{y - (\tilde{\alpha} + \mathbf{x}' \tilde{\beta})}{\sigma}\right)\right]. \quad (9.206)$$

9.3.6. Now let $\varepsilon = [y - E(y|\mathbf{x})]/\sigma$ be the scaled residual where $E(y|\mathbf{x}) = \tilde{\alpha} + \mathbf{x}' \tilde{\beta}$. Show that the density of ε is an extreme value (Gumbel) distribution with density $f(\varepsilon) = \exp[\varepsilon - e^\varepsilon]$. Thus, the log-linear model is of the form $y = \tilde{\alpha} + \mathbf{x}' \tilde{\beta} - \sigma \varepsilon$ with Gumbel distributed errors.

9.3.7. Now let $w = e^\varepsilon$. Show that the density function for w is the unit exponential, $f(w) = e^{-w}$.

9.3.8. The survival function is defined as $S(t) = P(T > t) = P[\log(T) > \log(t)]$. Since e^ε is distributed as the unit exponential, show that the survival function for a subject with covariate vector \mathbf{x} equals that presented in (9.204).

9.4 Log-Logistic Model. Consider the log-logistic distribution with rate parameter μ and shape parameter γ , where the hazard function is

$$\lambda(t) = \frac{\mu\gamma(t)^{\gamma-1}}{1 + \mu t^\gamma}. \quad (9.207)$$

9.4.1. Show that the survival function is $S(t) = [1 + \mu t^\gamma]^{-1}$.

9.4.2. Show that the survival distributions for two groups with the same shape parameter (γ) but with different rate parameters $\mu_1 \neq \mu_2$ have proportional failure and survival odds as in (9.51), with failure (event) odds ratio μ_1/μ_2 and survival odds ratio μ_2/μ_1 .

9.4.3. To generalize the log-logistic proportional odds model as a function of a covariate vector \mathbf{x} , let $\mu = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta})$. Show that

$$S(t|\mathbf{x}) = \left(1 + e^{\alpha + \mathbf{x}'\boldsymbol{\beta}} t^\gamma\right)^{-1}. \quad (9.208)$$

9.4.4. Show that the coefficient e^{β_j} equals the failure odds ratio per unit increase in the j th covariate x_j and that $e^{-\beta_j}$ equals the survival odds ratio.

9.4.5. Now, let $Y = \log(T)$. Show that the density of the distribution of $y = \log(t)$ is

$$f(y) = \frac{\mu e^{y\gamma}\gamma}{(1 + \mu e^{y\gamma})^2} \quad (9.209)$$

and the survival function is

$$S(y) = (1 + \mu e^{y\gamma})^{-1}. \quad (9.210)$$

9.4.6. Then for an individual with covariate vector \mathbf{x} , show that the conditional distribution is

$$f(y|\mathbf{x}) = \frac{\exp\left[\frac{y - (\tilde{\alpha} + \mathbf{x}'\tilde{\boldsymbol{\beta}})}{\sigma}\right]^{\frac{1}{\sigma}}}{\left[1 + \exp\left(\frac{y - (\tilde{\alpha} + \mathbf{x}'\tilde{\boldsymbol{\beta}})}{\sigma}\right)\right]^2} = \frac{\exp\left[\frac{y - E(y|\mathbf{x})}{\sigma}\right]^{\frac{1}{\sigma}}}{\left[1 + \exp\left(\frac{y - E(y|\mathbf{x})}{\sigma}\right)\right]^2}, \quad (9.211)$$

where the proportional odds model parameters are now expressed as

$$\mu = \exp[-E(y|\mathbf{x})\gamma] = \exp\left[-(\tilde{\alpha} + \mathbf{x}'\tilde{\boldsymbol{\beta}})/\sigma\right] \quad (9.212)$$

with $\gamma = \sigma^{-1}$, so that $\alpha = -\tilde{\alpha}/\sigma$, $\beta = -\tilde{\beta}/\sigma$. Therefore, the proportional failure or survival odds model parameters can be obtained from the accelerated failure-time model.

9.4.7. Then show that the survival function in $\log t$ can be expressed as

$$S(y|\mathbf{x}) = \left(1 + \exp \left[\frac{y - (\tilde{\alpha} + \mathbf{x}'\tilde{\beta})}{\sigma} \right] \right)^{-1} = \left(1 + \exp \left[\frac{y - E(y|\mathbf{x})}{\sigma} \right] \right)^{-1}. \quad (9.213)$$

9.4.8. Show that the density of $\varepsilon = [y - E(y|\mathbf{x})]/\sigma$ is the logistic distribution with density

$$f(\varepsilon) = \frac{e^\varepsilon}{[1 + e^\varepsilon]^2}. \quad (9.214)$$

Thus, the errors in the log-linear model are distributed as logistic and the hazard function in (9.207) corresponds to a log-logistic density.

9.5 Consider two populations with arbitrary hazard functions $\lambda_1(t)$ and $\lambda_2(t)$, respectively.

9.5.1. Show that the hazard ratio and the survival odds ratio cannot be proportional simultaneously under the same survival model. That is,

- a. If $\lambda_1(t)/\lambda_2(t) = \phi \forall t$, then $O_1(t)/O_2(t) \neq \text{constant } \forall t$.
- b. If $O_1(t)/O_2(t) = \varphi \forall t$, then $\lambda_1(t)/\lambda_2(t) \neq \text{constant } \forall t$, where $O_i(t) = S_i(t)/[1 - S_i(t)]$, $i = 1, 2$.

9.5.2. Show that two populations with Weibull survival distributions with rate parameters μ_1 and μ_2 , where $\mu_1 \neq \mu_2$ but with the same shape parameter $\gamma_1 = \gamma_2 = \gamma$ satisfy (a).

9.5.3. Show that two populations with logistic survival distributions with rate parameters μ_1 and μ_2 , where $\mu_1 \neq \mu_2$ but with the same shape parameter $\gamma_1 = \gamma_2 = \gamma$ satisfy (b).

9.6 Let T_a and T_b be two independent random variables corresponding to two separate causes of death, each with corresponding hazard functions $\lambda_a(t)$ and $\lambda_b(t)$, respectively. A subject will die from only one cause, and the observed death time for a subject is $t^* = \min(t_a, t_b)$, where t_a and t_b are the potential event times for that subject.

9.6.1. Show that t^* has the hazard function

$$\lambda^*(t^*) = \lambda_a(t^*) + \lambda_b(t^*). \quad (9.215)$$

Note that this generalizes to any number of competing causes of death.

9.6.2. Consider the case where one cause of death is the index event of interest with hazard function $\lambda_I(t)$, and the other is a competing risk with hazard $\lambda_C(t)$. Suppose that each is exponentially distributed with constant cause-specific hazards λ_I and λ_C . From (9.136) show that the subdistribution of the index event has density

$$f_I(t) = \lambda_I \exp [-(\lambda_I + \lambda_C) t]. \quad (9.216)$$

9.6.3. Now consider two populations with constant cause specific hazards $(\lambda_{I1}, \lambda_{C1})$ in the first population and $(\lambda_{I2}, \lambda_{C2})$ in the second. Show that the ratio of the ratio of the subdistribution densities is

$$\frac{f_{I1}(t)}{f_{I2}(t)} = \frac{\lambda_{I1}}{\lambda_{I2}} \exp [(\lambda_{I2} - \lambda_{I1} + \lambda_{C2} - \lambda_{C1})t]. \quad (9.217)$$

9.6.4. Show that if the cause-specific index hazards are the same in each population, $\lambda_{I1} = \lambda_{I2}$, but the competing risk hazards differ, then this index event density ratio is increasing (decreasing) in t when $\lambda_{C1} < (>) \lambda_{C2}$.

9.6.5. Use numerical examples to show that when the cause-specific index hazards differ, then depending on the values of the competing risk hazards, the above density ratio may remain constant over time, be increasing in t , or be decreasing in t . Also show that in the latter two cases the density ratio over time may cross from < 1 to > 1 , or vice versa.

9.6.6. For $\lambda_{I1} = 4, \lambda_{I2} = 2, \lambda_{C1} = 3, \lambda_{C2} = 1$, show that the index density ratio is decreasing in t and that the ratio equals 1 at $t = -\frac{1}{2} \log(1/2)$.

9.6.7. Section 9.2.3 presents an example involving competing risks where inclusion of competing risk exits in the risk set throughout yields a correct estimate of the survival probability. Show that this adjustment, in general, is appropriate when there are no exits other than those caused by the competing risk. For the j th interval, the extent of exposure or number at risk becomes $n_j = r_j + \sum_{\ell=1}^{j-1} e_{\ell} = N - \sum_{\ell=1}^{j-1} d_{\ell}$. Then show that the adjusted product-limit estimator at $t_{(j)}$ is

$$\widehat{S}(t_{(j)}) = \frac{N - \sum_{\ell=1}^j d_{\ell}}{N}. \quad (9.218)$$

9.7 Use the following set times of remission (in weeks) of leukemia subjects treated with 6-MP from Freireich et al. (1963), with censored survival times designated as t^+ (reproduced with permission):

6+ 6 6 6 7 9+ 10+ 10 11+ 13 16 17+ 19+ 20+ 22 23 25+ 32+ 32+ 34+ 35+

9.7.1. Assuming exponential survival, estimate λ and the variance of the estimate. Use $\widehat{\lambda}$ to estimate the survival function $\widehat{S}(t)$ and its 95% asymmetric confidence bands based on the estimated variance of $\log[\widehat{S}(t)]$.

9.7.2. Alternatively, assuming a Weibull distribution, use Newton-Raphson iteration or SAS PROC LIFEREG to compute the maximum likelihood estimates of the parameters (μ, γ) and the estimated covariance matrix of the estimates. Note that PROC LIFEREG uses the extreme value representation in Problem 9.3 to provide estimates of α and σ^2 in a no-covariate model. Use these to compute the estimated hazard function, the estimated survival function, and the asymmetric 95% confidence bands.

9.7.3. Likewise, assuming a log-logistic distribution, use Newton-Raphson iteration or SAS PROC LIFEREG to compute the maximum likelihood estimates of the

parameters (μ, γ) and the estimated covariance matrix of the estimates. Use these to compute the estimated hazard function, the estimated survival function, and the asymmetric 95% confidence bands.

9.7.4. Compute the Kaplan-Meier product limit estimate of the survival function and its asymmetric 95% confidence band.

9.7.5. Is there any substantial difference among these estimates of the survival distribution and its confidence bands?

9.8 Consider the case of a discrete survival distribution with lumps of probability mass of the event f_j at the j th discrete event time $t_{(j)}$, $1 \leq j \leq J$. Then the event probability π_j at the j th event time is defined as

$$\pi_j = \frac{f_j}{\sum_{\ell \geq j} f_\ell}. \quad (9.219)$$

9.8.1. Show that

$$S(t_{(j)}) = \prod_{\ell \leq j} [1 - \pi_\ell] = 1 - \sum_{\ell \leq j} f_\ell. \quad (9.220)$$

9.8.2. At the j th discrete event time $t_{(j)}$, let n_j be the number at risk, and let d_j be the number of events observed, $d_j > 0$. Also let w_j be the number of observations censored during the interval $(t_{(j)}, t_{(j+1)})$ between the j th and $(j+1)$ th event times. Then $n_{j+1} = n_j - d_j - w_j$. Using the product representation of the survival function in (9.220), show that

$$\prod_{j=1}^J S(t_{(j-1)})^{d_j} S(t_{(j)})^{w_j} = \prod_{j=1}^J (1 - \pi_j)^{n_j - d_j}, \quad (9.221)$$

thus deriving the simplification in (9.14) from the likelihood in (9.10).

9.8.3. Using similar steps, show that the modified likelihood in (9.39) yields the actuarial estimate $\hat{\pi}_j = p_j$ in (9.40).

9.8.4. The product-limit or actuarial estimator of the survivor function can be expressed as $\hat{S}(t_{(j)}) = \prod_{\ell=1}^j q_\ell$, as shown in (9.18) and (9.42), respectively, where the estimate of the continuation probability is $q_j = 1 - p_j$ and $p_j = d_j/n_j$. Consider the estimates p_j and p_{j+1} at two successive event times $t_{(j)}$ and $t_{(j+1)}$. Despite the clear dependence of the denominator n_{j+1} of p_{j+1} on the elements of p_j , show that p_j and p_{j+1} are uncorrelated. Do likewise for any pair p_j and p_k for $1 \leq j < k \leq J$. Thus, asymptotically the large sample distribution of the vector of proportions $\mathbf{p} = (p_1 \cdots p_J)^T$ is asymptotically distributed as multivariate normal with expectation $\boldsymbol{\pi} = (\pi_1 \cdots \pi_J)^T$ and covariance matrix $\boldsymbol{\Sigma}$ that is diagonal with elements $\sigma_j^2 = \pi_j(1 - \pi_j)/n_j$.

9.8.5. Then using the δ -method, derive the variance of $\log[\hat{S}(t_{(j)})]$ presented in (9.20).

9.8.6. Again use the δ -method to derive Greenwood's expression for the variance of $\hat{S}(t_{(j)})$ shown in (9.22). This yields the large sample S.E. of $\hat{S}(t)$ and a confidence interval which is symmetric about $\hat{S}(t)$, but not bounded by 0 and 1.

9.8.7. To obtain a bounded confidence interval, derive the expression for the variance of the $\log(-\log[\widehat{S}(t)])$ in (9.23).

9.8.8. Noting that $-\log[\widehat{S}(t)] = \widehat{\Lambda}(t)$, derive the asymmetric confidence limits presented in (9.24).

9.8.9. As an alternative to the complementary log-log transformation, show that the variance of the logit of the survival probability at time t is

$$\widehat{V} \left[\log \left(\frac{\widehat{S}(t)}{1 - \widehat{S}(t)} \right) \right] = \left(\frac{1}{1 - \widehat{S}(t)} \right)^2 \widehat{V} \left(\log [\widehat{S}(t)] \right). \quad (9.222)$$

This also yields asymmetric confidence limits on the survival odds ratio at time t . Then use the logistic function to obtain the expression for the asymmetric confidence limits on the survival probability $S(t)$.

9.8.10. Show that the Kaplan-Meier estimate of the survival function estimate implies an estimate of the hazard function at time $t_{(j)}$, which is obtained as

$$\widehat{\lambda}_{KM,j} = \frac{-\log q_j}{t_{(j)} - t_{(j-1)}} = \frac{-\log(1 - p_j)}{t_{(j)} - t_{(j-1)}}. \quad (9.223)$$

This is also called Peterson's estimate (Peterson, 1977).

9.8.11. Asymptotically as $J \rightarrow \infty$ and $p_j \downarrow 0$, show that $\widehat{\lambda}_{KM,j} \cong \widehat{\lambda}_j$ in (9.26) and that the Kaplan-Meier estimate of the survival function $\widehat{S}(t)$ in (9.18) is approximately equal to the Nelson-Aalen estimate $\widehat{S}_{NA}(t)$ in (9.28). *Hint:* Use $\Lambda_{NA}(t_{(j)})$ and note that $\log(1 - \epsilon) \doteq \epsilon$ for $\epsilon \downarrow 0$.

9.9 Consider a sample of N observations where the j th subject has a maximum exposure time e_j at which that subject's follow-up is right censored if the event has not been observed to occur. Then the indicator variable δ_j denotes whether the observed time t_j is an event time ($\delta_j = 1$) or a censoring time ($\delta_j = 0$). The overall proportion of events then is $p = \sum_j \delta_j / N$. The naive binomial variance is $V(p) = \pi(1 - \pi)/N$, where $\pi = E(p)$.

9.9.1. Condition on the set of exposure times (e_1, \dots, e_N) as fixed quantities. Then, given the underlying cumulative distribution and survival functions $F(t)$ and $S(t)$, respectively, show that

$$E(p) = \pi = \frac{\sum_j E(\delta_j | e_j)}{N} = \frac{\sum_j F(e_j)}{N}. \quad (9.224)$$

9.9.2. Also show that the variance of p is

$$V_c(p) = \frac{\sum_j V(\delta_j | e_j)}{N} = \frac{\sum_j F(e_j)S(e_j)}{N}. \quad (9.225)$$

9.9.3. Show that this correct variance $V_c(p) \leq V(p)$, where $V(p)$ is the naive binomial variance. *Hint:* Note that $V(\delta) = E[V(\delta | t)] + V[E(\delta | t)]$.

9.10 The proportional odds model assumes that the survival odds over time for a subject with covariate vector \boldsymbol{x} is proportional to the background survival odds.

Bennett (1983), among others, describes the proportional odds model where the survival function is of the form

$$S(t|\mathbf{x}) = \left[1 + e^{\alpha_0(t) + \mathbf{x}'\boldsymbol{\beta}} \right]^{-1}, \quad (9.226)$$

where $\alpha_0(t)$ is some function of time that is expressed as a function of additional parameters.

9.10.1. Show that the survival odds for two subjects with covariate values \mathbf{x}_1 and \mathbf{x}_2 are proportional over time.

9.10.2. Show that the cumulative distribution is a logistic function of $\alpha_0(t) + \mathbf{x}'\boldsymbol{\beta}$.

9.10.3. Show that the hazard ratio for two subjects with covariate values \mathbf{x}_1 and \mathbf{x}_2 is

$$\frac{\lambda(t|\mathbf{x}_2)}{\lambda(t|\mathbf{x}_1)} = \frac{F(t|\mathbf{x}_2)}{F(t|\mathbf{x}_1)}. \quad (9.227)$$

9.10.4. Now consider a single binary indicator variable to represent treatment group, $x = (0, 1)$, with coefficient β . Then let

$$H(\beta) = \log \frac{\lambda(t|x=1)}{\lambda(t|x=0)}. \quad (9.228)$$

Under the null hypothesis of no treatment group difference, $H_0: \beta = \beta_0 = 0$, use a Taylor's expansion of $H(\beta)$ about $H(\beta_0)$ to show that

$$H(\beta) \cong (\beta - \beta_0)H'(\beta_0). \quad (9.229)$$

9.10.5. Then show that

$$H'(\beta) = 1 - F(t|x). \quad (9.230)$$

Evaluating $H'(\beta)$ under the null hypothesis at $\beta = \beta_0$, it follows that

$$\log \frac{\lambda(t|x=1)}{\lambda(t|x=0)} \cong (\beta - \beta_0)[1 - F_0(t)] = g[F_0(t)], \quad (9.231)$$

which is a function of $F_0(t)$, thus satisfying Schoenfeld's (1981) condition in (9.46).

9.10.6. Using (9.47), show that the asymptotically efficient test against the alternative of proportional odds over time for two groups is the weighted Mantel-Haenszel test in (9.43) with weights $w(t) = \widehat{S}(t)$ based on the Kaplan-Meier estimate of the survival function in the combined sample. This is the Peto-Peto-Prentice Wilcoxon test.

9.11 The following table presents the data from Freireich et al. (1963) with the times of remission in weeks of leukemia subjects treated with placebo:

1 1 2 2 3 4 4 5 5 8 8 8 11 11 12 12 15 17 22 23

where no event times were censored. We now wish to compare the disease-free survival (remission) times between the group of patients treated with 6-MP presented in Problem 9.7 with those treated with placebo.

9.11.1. Order the distinct observed death times (ignoring ties or multiplicities) as $(t_{(1)}, \dots, t_{(J)})$ for the J distinct times ($J \leq d_* = \#$ events) in the combined sample.

9.11.2. Compute the number at risk n_{ij} in the i th group ($i = 1$ for 6-MP, 2 for placebo) and the number of events d_{ij} at the j th remission time $t_{(j)}$, $j = 1, \dots, J$. Construct a table with columns $t_{(j)}$, d_j , N_j , d_{1j} , n_{1j} , d_{2j} , n_{2j} . Note that for each event time we can construct a 2×2 table of treatment (1 or 2) versus event (remission or not), with N_j being the total sample size for that table.

9.11.3. From these data, calculate the Kaplan-Meier estimator of the survivor distribution and its standard errors for each of the treatment groups.

9.11.4. Compute the $S.E.$ of the $\log(-\log[\hat{S}(t)])$ and the 95% asymmetric confidence limits for each survivor function.

9.11.5. Likewise, using the expressions in Section 9.1.4, compute the piecewise constant (linearized) hazard function estimate $\hat{\lambda}_{(j)}$, the Nelson-Aalen estimate of the cumulative hazard $\hat{\Lambda}_{NA}(t)$, and the Nelson-Aalen estimate of the survivor function $\hat{S}_{NA}(t)$ for each group.

9.11.6. At each event time, compute $\hat{V}[\hat{\Lambda}_{NA}(t_{(j)})]$ and the corresponding $S.E.$ From these, for each group compute the asymmetric 95% confidence limits on $S(t)$ obtained via confidence limits on $\Lambda(t)$.

9.11.7. Compute the Mantel-logrank, Gehan-Wilcoxon, and the Peto-Peto-Prentice Wilcoxon tests of equality of the hazard functions. In practice, the specific test to be used would be specified a priori. From examination of the linearized hazard function estimates, explain why the logrank test yields the larger test value.

9.11.8. As described in Section 9.3.5, compute the Peto approximate estimate of the relative risk and its 95% confidence limits, and also for the survival odds.

9.11.9. Using the methods described in Section 4.9.2, compute the correlation between the Mantel-logrank and the Peto-Peto-Prentice Wilcoxon tests, the MERT combination of these two tests, and the estimated maximin efficiency of the test.

9.12 Table 9.7 presents data showing the incidence of cardiovascular mortality among subjects treated with tolbutamide versus placebo over eight years of treatment and follow-up in the University Group Diabetes Program (UGDP, 1970). This is an example of grouped-time survival data.

9.12.1. Compute the actuarial adjusted number of units of exposure at risk n_j , n_{1j} , and n_{2j} during each interval, and from these calculate the actuarial estimator of the survivor function for each treatment group and their standard errors using Greenwood's equation.

9.12.2. Compute the logrank test, the Gehan-Wilcoxon test, the Peto-Peto-Prentice Wilcoxon test, and the MERT to compare the two groups. In practice the test to be used should be specified a priori.

9.12.3. Compare the results of these tests based on the nature of the differences between groups with respect to the survival distributions and the censoring distributions.

Table 9.7 Number entering each year of follow-up, numbers withdrawn alive during that year, and number of deaths during the year for the tolbutamide and placebo groups of the UGDP (reproduced with permission).

Year	Placebo			Tolbutamide		
	n_{1j}	w_{1j}	d_{1j}	n_{2j}	w_{2j}	d_{2j}
1	205	0	0	204	0	0
2	205	0	5	204	0	5
3	200	0	4	199	0	5
4	196	4	4	194	5	5
5	188	23	4	184	24	5
6	161	43	3	155	41	4
7	115	50	1	110	47	5
8	64	36	0	58	33	1

9.13 Table 9.8 presents observations from a hypothetical time-to-event study in discrete time. Here we studied an experimental drug versus placebo for treatment of bone cancer in an extremity among an elderly population. Patients were *x*-rayed every two months after treatment was started to see if the cancer had gone into remission (i.e., the subject had healed). The outcome for each subject was categorized as either healed at a designated evaluation visit at week 2, 4, or 6; the subject died during the stated month (and prior *x*-rays showed that the subject had not yet healed); the limb was removed surgically during the stated month because the cancer had spread; the subject dropped out (withdrew) during the stated month; or the subject had not gone into remission (healed) after completing the six months of follow-up. For each subject, the month of the last *x*-ray is also given.

9.13.1. Using the standard intervals (0–2] months, (2–4] months, and (4–6] months; construct a table that shows the timing of losses to follow-up in relation to the time of the last examination. As described in Section 9.2.3, construct a work-table that shows the numbers evaluated at each time, the number healed, and the numbers "censored" between evaluations.

9.13.2. Censoring on Death/Surgery. Treat death, surgery, and lost-to-follow-up as censored at random immediately following the last *x*-ray evaluation. From this, present a work table with the numbers of events and the adjusted numbers at risk separately for each group and for both groups combined. Within each treatment group, compute the modified Kaplan-Meier "survival" or "disease duration" function and the linearized estimate of the hazard function. Also compute the Mantel-Logrank and Peto-Peto-Prentice Wilcoxon tests. Note that this survival function estimates the probability of continuing to have cancer in a population where no one dies or has limb amputation. The hazard function is an estimate of the cause-specific hazard function that also forms the basis for the statistical tests.

9.13.3. Competing Risk Adjustment. Treat death and surgery as competing risks that are not censored during the study but are retained in the adjusted numbers

Table 9.8 Data from a hypothetical clinical trial of drug treatment for bone cancer.

Drug Group			Placebo Group		
Status	Month	Last x-ray	Status	Month	Last x-ray
Healed	2		Healed	2	
Healed	2		Healed	2	
Healed	2		Lost	1	0
Healed	2		Lost	1	0
Healed	2		Lost	2	0
Healed	2		Surgery	2	0
Lost	1	0	Surgery	2	0
Lost	2	0	Death	1	0
Surgery	1	0	Healed	4	
Healed	4	0	Healed	4	
Healed	4		Lost	3	2
Healed	4		Lost	4	2
Healed	4		Surgery	4	2
Death	3	2	Surgery	3	2
Lost	3	2	Surgery	4	2
Lost	4	2	Healed	6	
Lost	4	2	Healed	6	
Surgery	3	2	Lost	5	4
Surgery	4	2	Lost	5	4
Healed	6		Lost	5	4
Healed	6		Surgery	6	4
Lost	5	4	Surgery	5	4
Lost	5	4	Death	5	4
Surgery	6	4	Not Healed	6	6
Not Healed	6	6			
Not Healed	6	6			
Not Healed	6	6			
Not Healed	6	6			

at risk. Also treat lost-to-follow-up as censoring at random. From this, present a work table with the numbers of events and the adjusted numbers at risk separately for each group. Use this table to calculate the modified Kaplan-Meier intent-to-treat survival function, including the standard errors of the survival function, separately for each treatment group. Also compute the Mantel-logrank and Peto-Peto-Prentice Wilcoxon tests.

9.13.4. Sub-Distribution Function. Now use a combined outcome of healing, death, or surgery and treat losses-to-follow-up as censored after the last x-ray evaluation. Compute the estimate of the survival function for the combined outcome, designated as $\widehat{S}_{I,C}(t)$ in Section 9.2.3. Then calculate the cumulative subdistribution function for healing $\widehat{F}_I(t)$ using (9.138). Note that the estimates of $\widehat{F}_I(t)$ are approximately equal to $1 - \widehat{S}(t)$ in Problem 9.13.3. In the drug-treated group, the values of $\widehat{F}_I(t)$ at 2, 4, and 6 months are 0.231, 0.413, and 0.531, respectively, whereas the estimates of $\widehat{S}(t)$ in Problem 9.3.3 are 0.769, 0.588, and 0.481, respectively. For a definitive analysis, the subdistribution function calculation is preferred. In this case, the standard errors would be calculated using the results of Gaynor et al. (1993) and the tests of significance computed as described by Gray (1988). These computations are described by Marubini and Valsecchi (1995).

9.14 Consider the case of J distinct event times $t_{(1)} < t_{(2)} < \dots < t_{(J)}$, where allowing for ties, the total number events at $t_{(j)}$ is $d_j \geq 1$ for all j , $1 \leq j \leq J$, where $d_j > 1$ indicates tied event times.

9.14.1. Using the Peto-Breslow approximate likelihood allowing for ties in (9.99), show that the k th element of the score vector $\mathbf{U}(\boldsymbol{\beta})$ corresponding to coefficient β_k is

$$\mathbf{U}(\boldsymbol{\beta})_{\beta_k} = \frac{\partial \log \widetilde{L}_{PB}(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{j=1}^J s_{jk} - \bar{x}_k(t_j, \boldsymbol{\beta}), \quad (9.232)$$

where $s_{jk} = \sum_{j=1}^{d_j} x_{jk}$ for the d_j subjects with the event at time t_j , and where $\bar{x}_k(t_j, \boldsymbol{\beta})$ is as shown in (9.78) with the subscript i changed to j to designate the j th event time.

9.14.2. Then show that the information matrix has elements

$$\mathbf{I}(\boldsymbol{\beta})_{km} = E \left[\frac{-\partial \log \widetilde{L}_{PB}(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_m} \right] = \sum_{j=1}^J [C_{jkm}(t_j, \boldsymbol{\beta}) - \bar{x}_k(t_j, \boldsymbol{\beta}) \bar{x}_m(t_j, \boldsymbol{\beta})] \quad (9.233)$$

with elements $C_{jkm}(t_j, \boldsymbol{\beta})$ as shown in (9.80) evaluated at the j th event time t_j for $1 \leq k \leq m \leq p$.

9.14.3. Now consider that we wish to use a Cox proportional hazards model with a solitary binary covariate $x_i = 1$ or 0 for $i = 1, \dots, N$, such as where X designates treatment or exposure group. Using the Peto-Breslow approximate likelihood allowing for ties, show that the score equation for $\boldsymbol{\beta}$ can be expressed as

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{j=1}^J d_{1j} - \frac{m_{1j} n_{1j} e^{\boldsymbol{\beta}}}{n_{1j} e^{\boldsymbol{\beta}} + n_{2j}}, \quad (9.234)$$

where at the j th event time $t_{(j)}$, d_{ij} is the number of events among the n_{ij} subjects at risk in the i th group and m_{1j} is the total number of events in both groups combined, $m_{1j} = d_j = d_{1j} + d_{2j}$. Thus, under the null hypothesis $H_0: \beta = \beta_0 = 0$, the score is

$$U(\beta_0) = \sum_{j=1}^J d_{1j} - \frac{m_{1j}n_{1j}}{N_j} = \sum_{j=1}^J d_{1j} - E(d_{1j}|H_0), \quad (9.235)$$

where $N_j = n_{1j} + n_{2j}$.

9.14.4. Show that the expected information then is

$$I(\beta) = \sum_{j=1}^J \frac{m_{1j}n_{1j}e^\beta n_{2j}}{(n_{1j}e^\beta + n_{2j})^2}, \quad (9.236)$$

which, when evaluated under the null hypothesis, yields

$$I(\beta_0) = \sum_{j=1}^J \frac{m_{1j}n_{1j}n_{2j}}{N_j^2}. \quad (9.237)$$

9.14.5. Then show that the score test equals the Mantel-logrank test with no ties. For the case of ties, show that the efficient score test variance $I(\beta_0)$ differs from the Mantel-logrank test variance in (9.43) by the factor $m_{2j}/(N_j - 1)$, where $m_{2j} = N_j - m_{1j}$.

9.14.6. Using the score equation for the PH model with no ties, use a Taylor's expansion about the null value as described in Section 6.4 to show that the Peto score-based estimate of the log relative risk equals that described in Section 9.3.5 based on the logrank test.

9.14.7. Now consider the case of K independent strata, where it is assumed that there is a constant ratio of the hazards of the two groups over time within all strata. This entails the stratified likelihood in (9.73), where $\beta_h = \beta$ for all strata but where the background hazard functions within each stratum may be different. With no ties, generalize the above results to show that the efficient stratified-adjusted score test equals the stratified-adjusted Mantel-logrank test presented in (9.52).

9.14.8. To allow for ties, now use the Cox conditional likelihood in (9.98). Show that the efficient scores test for $H_0: \beta = 0$ equals the Mantel-logrank test in (9.43) with $w_j = 1$. In this derivation, note that the score equation and the information are functions of the mean and variance, respectively, of a hypergeometric random variable, from which the result follows.

9.14.9. Now consider a model with multiple covariates. If one of the covariates, say X_1 , is a binary indicator variable for group membership, 1 if experimental, 0 if control, then show that the hazard ratio $\phi = \lambda_e(t)/\lambda_c(t) = e^{\beta_1}$.

9.14.10. Alternatively, let X_1 be a quantitative variable. Show that $e^{\beta_1} = \lambda(t|x+1)/\lambda(t|x)$ is the relative hazard per unit increase in the value of the covariate, the values of other covariates held constant.

9.14.11. Now let X_1 be the log of a quantitative variable, $X_1 = \log(Z_1)$. Consider two subjects, say i and j , with values x_{i1} and x_{j1} (z_{i1} and z_{j1}). Express the relative hazards for these two subjects in terms of the value of β_1 , the values of other

covariates held constant. Show that if $z_{i1}/z_{j1} = c$ for $c > 0$, then the relative hazard for these two subjects is c^{β_1} . Thus, the relative percentage change in the hazard is $100(c^{\beta_1} - 1)$ per c -fold change in Z .

9.14.12. The discrete- or grouped-time model of Section 9.4.6.2 assumes that a logistic model applies to the j th interval such that

$$\frac{\pi_{j|\mathbf{x}}}{1 - \pi_{j|\mathbf{x}}} = \frac{\pi_{0j}}{1 - \pi_{0j}} e^{\mathbf{x}'\boldsymbol{\beta}} \quad (9.238)$$

for $1 < j < K$, where $\pi_{0j} = \pi_{j|\mathbf{x}=0} = 1 - \int_{\tau_{j-1}}^{\tau_j} \lambda_0(t) dt$ and $\pi_{j|\mathbf{x}} = 1 - \int_{\tau_{j-1}}^{\tau_j} \lambda(t|\mathbf{x}) dt$. Show that as $K \rightarrow \infty$, such that the $\tau_j \downarrow \tau_{j-1} \forall j$, this model approaches the continuous-time PH model in (9.57). *Hint:* Use L'Hospital's rule.

9.15 Consider planning a two-group clinical trial to test the hypothesis $H_0: \lambda_1 = \lambda_2$ under exponential model assumptions at $\alpha = 0.05$, two-sided. We desire 90% power to detect a specific hazard ratio $\phi = \lambda_1/\lambda_2$ for a specified hazard rate in the control group λ_2 .

9.15.1. Assume no censoring other than administrative censoring with uniform entry over the interval $(0, T_R]$ and total study duration $T_S > T_R$. Consider each of the following designs:

Design	λ_2	ϕ	T_R	T_S
1	0.1	2	3	5
2	0.1	2	5	5
3	0.2	2	3	5
4	0.1	2.5	3	5

Assuming equal allocations to the two groups, compute the required total sample N , the expected proportion of events within each group $E(\delta|\lambda_1)$ and $E(\delta|\lambda_2)$, and that under the null hypothesis $E(\delta|\lambda)$.

9.15.2. What is the implication of the difference between designs 1 and 2 ($T_R = 3$ vs. 5 years)?

9.15.3. What is the implication of the difference between designs 1 and 3 ($\lambda_2 = 0.1$ vs. 0.2)?

9.15.4. What is the implication of the difference between designs 1 and 4 ($\phi = 2$ vs. 2.5)?

9.15.5. Now assume random censoring with an exponential hazard rate $\eta = 0.05$ within each of the two groups. For each of the four designs, compute the above quantities plus the expected proportions lost to follow-up within each group $E(\xi|\lambda_1, \eta)$ and $E(\xi|\lambda_2, \eta)$ and that under the null hypothesis $E(\xi|\lambda, \eta)$.

9.15.6. What impact do these losses have on the relative comparisons in Problems 9.15.2, 9.15.3, and 9.15.4?

9.16 Table 9.9 presents data from Prentice (1973) for subjects with lung cancer treated by a standard or by a test method. The covariates are Z_1 : performance status with nine levels that is treated as a quantitative covariate; Z_2 : time since diagnosis;

Table 9.9 Prentice (1973) lung cancer data, reproduced with permission.

<i>ID</i>	<i>t_i</i>	δ_i	<i>Z₁</i>	<i>Z₂</i>	<i>Z₃</i>	<i>Z₄</i>	<i>Z₅</i>	<i>ID</i>	<i>t_i</i>	δ_i	<i>Z₁</i>	<i>Z₂</i>	<i>Z₃</i>	<i>Z₄</i>	<i>Z₅</i>
16	999	1	90	12	54	1	1	8	110	1	80	29	68	0	0
21	991	1	70	7	50	1	1	10	100	0	70	6	70	0	0
24	587	1	60	3	58	0	1	18	87	0	80	3	48	0	1
29	457	1	90	2	64	0	1	7	82	1	40	10	69	1	0
2	411	1	70	5	64	1	0	1	72	1	60	7	69	0	0
25	389	1	90	2	62	0	1	33	44	1	60	13	70	1	1
28	357	1	70	13	58	0	1	11	42	1	60	4	81	0	0
9	314	1	50	18	43	0	0	26	33	1	30	6	64	0	1
34	283	1	90	2	51	0	1	32	30	1	70	11	63	0	1
20	242	1	50	1	70	0	1	27	25	1	20	36	63	0	1
19	231	0	50	8	52	1	1	14	25	0	80	9	52	1	0
3	228	1	60	3	38	0	0	35	15	1	50	13	40	1	1
30	201	1	80	28	52	1	1	15	11	1	70	11	48	1	0
13	141	1	30	4	63	0	0	6	10	1	10	5	49	0	0
4	126	1	60	9	63	1	0	12	8	1	40	58	63	1	0
5	118	1	70	11	65	1	0	23	1	1	20	21	65	1	1
17	112	1	80	6	60	0	1	31	1	1	50	7	35	0	1
22	111	1	70	3	62	0	1								

Z_3 : age (years) at diagnosis; Z_4 : any prior treatment (1 = yes, 0 = no); and Z_5 : study drug (1 = test, 0 = standard). Survival times are recorded to the nearest day since entry into the study and $\delta_i = I(\text{death})$.

The principal objectives of this analysis are to compare the study drug treatment groups (test vs. standard) to assess the impact of other covariates on survival and to assess the differences between treatment groups, adjusting for these other covariates.

9.16.1. Compute and plot the Kaplan-Meier estimated survival functions and the Nelson-Aalen estimates of the cumulative hazard functions for the two drug groups (Z_5).

9.16.2. Compute the unadjusted logrank test comparing the drug groups and an estimate of the average relative risk (hazard ratio) using the Peto estimate along with its 95% confidence limits.

9.16.3. Compute an estimate of the average relative risk (hazard ratio) from an unadjusted PH model and its 95% confidence limits.

9.16.4. Compute a measure of explained variation using Madalla's R_{LR}^2 in (A.210) and also using $R_{\varepsilon^2}^2$ in (9.108) with $\sigma_{\varepsilon}^2 = 1$; i.e., using the Kent and O'Quigley $R_{W,A}^2$.

9.16.5. The covariate Z_4 is binary. Consider the model where Z_4 is used to define strata and where the covariates within each strata are $Z = (Z_1, Z_2, Z_3, Z_5)$. Compute $\log[-\log[\hat{S}_j(t)]]$ plots for the two strata ($j = 1, 2$) to assess whether the PH assumption for Z_4 is appropriate.

- 9.16.6.** Do likewise for Z_5 used to define strata and for (Z_1, Z_2, Z_3, Z_4) .
- 9.16.7.** Fit a PH model with an interaction between Z_4 and $\log(t)$ to assess whether a monotone departure in $\log t$ from the PH assumption exists for this covariate.
- 9.16.8.** Likewise, fit a PH model with an interaction between Z_5 and $\log(t)$ to assess whether a monotone departure from the PH assumption exists for the differences between drug treatment groups. Since there are no substantial departures from the PH model assumptions for these two covariates, there is no need to adopt remedial measures.
- 9.16.9.** Now fit a model with the adjusting covariates only (Z_1, Z_2, Z_3, Z_4) . Use $R^2_{\varepsilon^2}$ in (9.108) to provide a measure of explained variation for the model.
- 9.16.10.** For an "average" individual with covariate vector equal to the mean values for these adjusting covariates $\mathbf{x} = \bar{\mathbf{x}}$ (including the binary covariate Z_4), describe the estimated survival function $\hat{S}(t|\bar{\mathbf{x}})$.
- 9.16.11.** The coefficients are negative for (Z_1, Z_3) and positive for (Z_2, Z_4) . Thus, consider a high-risk subject with covariate values equal to the 25th, 75th, 75th and 25th percentiles for (Z_1, Z_2, Z_3, Z_4) , respectively; and a low-risk subject with values equal to the 75th, 25th, 25th and 75th percentiles. Compute and plot the estimated survival function $\hat{S}(t|\mathbf{x})$ for each of these two patients.
- 9.16.12.** Now fit a model with treatment group adjusting for the other covariates (Z_1, Z_2, Z_3, Z_4) . Compute the likelihood ratio test of the treatment group effect from the change in the model chi-square compared to the covariate-only model. Use Madalla's approximate R^2_{LR} based on this chi-square test value to describe the partial variation explained by treatment group. Also compute the $R^2_{W,A}$ in (9.109) to describe the treatment group effect.
- 9.16.13.** Then fit a model with a simple interaction between treatment group and each of the other covariates to assess the presence of interactions in the regression coefficients.
- 9.16.14.** Also fit a model stratifying on the value of Z_4 with no interactions with other covariates but with separate treatment group coefficients within each Z_4 stratum. Conduct a Wald test of the difference between the treatment group coefficients within the two strata. Because there is no indication of a treatment group by covariate (or by stratum) interaction, the no-interaction adjusted model appears adequate.
- 9.16.15.** Using the no-interaction model, compute the information sandwich robust estimate of the covariance matrix of the coefficient estimates. Also compute the robust model score test for the model on 5 df . Compared to the model-based covariance of the estimates and score test, does it appear that there is gross model misspecification?
- 9.17** Byar (1985) presents the data from the Veterans Administration Cooperative Urologic Research Group (VACURG) study of the treatment of prostate cancer with high- versus low-dose estrogen versus placebo. These data are available from StatLib. Analyses of these data were presented by Byar (1985) and Thall and Lachin (1986), among others. For this problem, some manipulations are required, the SAS program

for which is available online (see the Preface). Here we employ only the data from the high- versus low-dose groups. The variables in this data set are *pat* = patient number (not a covariate); *trt* = 1 (high dose), 0 (low dose) estrogen; *died* = 1 (died), 0 (survived, censored); *mosfu* = number of months of follow-up to death or censoring; *history* = 1 (yes), 0 (no) for history of cardiovascular disease; *age* in years; and *size* (of tumor) = 1 (large), 0 (small). The objective is to examine covariate effects on the risk of all-cause mortality, including cardiovascular mortality, for which a history of cardiovascular disease and age are the dominant risk factors. Perform the following analyses of these data.

9.17.1. Compute a Kaplan-Meier estimate of the survival function within each treatment group and conduct a logrank test of the difference between groups. Also compute the score-based estimate of the unadjusted relative risk and its 95% confidence limits.

9.17.2. Fit a PH model to obtain an unadjusted estimate of the relative risk and its 95% confidence limits.

9.17.3. Fit a PH model with *trt*, *history*, *age*, and *size* as covariates to obtain an adjusted estimate of the relative risk for treatment group. Interpret each coefficient in terms of the hazard ratio (relative risk) with asymmetric confidence limits.

9.17.4. For *age*, also compute the relative risk per ten years of age and the associated confidence limits.

9.17.5. Also, fit a model without *trt* and compute a likelihood ratio test of the effect of treatment.

9.17.6. State your conclusion as to the effects of high-dose estrogens on all-cause mortality.

9.17.7. Compute the robust information sandwich estimator of the covariance matrix of the estimates and the associated robust 95% confidence limits and the robust Wald tests for each covariate. Compare these to those obtained under the assumed model.

9.18 Problem 8.10 describes the data presented by Fleming and Harrington (1991) on the rate of serious infections among children in a clinical trial of interferon versus placebo. There, only the total number of infections experienced by each child were considered. Here, we consider the actual times of infection and the underlying intensity functions over time. The more complete data set includes one record per interval of time per subject. Each interval is designated by a start time (T_2) and a stop time (T_1), where T_1 is either the time of an event designated by $d = 1$ or the time of curtailment of follow-up or right censoring ($d = 0$). The record number for each subject is designated by the variable *s*. The complete event-time data are also available online (see the Preface, reproduced with permission).

9.18.1. Separately within each treatment group, compute the Nelson-Aalen estimate of the cumulative intensity function over time.

9.18.2. Also, compute the linearized estimate of the underlying intensity function, or the kernel-smoothed estimate using Epanechnikov's kernel.

9.18.3. Compute the Aalen-Gill logrank test of the difference between intensities using all recurrent events.

9.18.4. Fit the multiplicative intensity model with treatment group as the only covariate to obtain an unadjusted estimate of the relative risk and its confidence limits. Compare the Wald and likelihood ratio tests for the treatment group effect to the logrank Aalen-Gill test.

9.18.5. Fit a more complete model to assess the treatment group effect adjusted for the other covariates. Note that three separate binary indicator variables must be defined to reflect the effect of hospital type. Compare the adjusted to the unadjusted assessment of the treatment effect.

9.18.6. Fit an additional model that includes an interaction between treatment group and weight and an interaction between group and height. Describe the effect of treatment as a function of height and weight.

Appendix

Statistical Theory

A.1 INTRODUCTION

This appendix contains all of the mathematical statistical developments employed in the text. It is intended to be used hand-in-hand with the chapters in the main body of the book. In keeping with the general tone of the text, these developments are presented in an informal but precise manner. Readers who desire a more rigorous development of these results, such as the required regularity conditions, are referred to one of the many excellent texts on theoretical statistics, such as those by Bickel and Doksum (1977), Cox and Hinkley (1974), Cramér (1946), Kendall and Stuart (1979), Lehmann (1983, 1986), and Rao (1973).

A.1.1 Notation

Throughout the book, the following notation conventions are employed.

$E(Y)$ and $V(Y)$ denote the mean and variance of the random variable Y .

n and N denote the fixed total sample size in a single group or all total, respectively.

n is also used to denote an increasing sequence of n observations as $n \rightarrow \infty$.

$\{y_i\}$ denotes the set of values, in this case the set (y_1, \dots, y_N) of size N .

The notation \xrightarrow{p} refers to convergence in probability.

The notation \xrightarrow{d} refers to convergence in distribution, or in law.

The symbol $\hat{=}$ means "estimated as" as in $\mu \hat{=} \bar{y}$.

The symbol \sim means "is distributed as."

The symbol \cong means "is asymptotically equal to."

The symbol \doteq means "is approximately equal to."

The symbol \approx means "is asymptotically distributed as."

The symbol $|$ is used as a conditioning operator as in $E(y|x)$.

The notation $f(y; \theta)$ is used to designate the distribution of Y with parameter vector θ .

$I\{\text{expression}\}$ is used to designate the indicator function that equals 1 if the expression is true, 0 if false.

A.1.2 Matrices

A review of matrix algebra will not be provided. Selected results will be presented as needed. The following conventions are employed.

All vectors are column vectors, such that $\theta = (\theta_1 \ \dots \ \theta_p)^T$, the " T " being the transpose operator. Alternatively, θ' is used to designate the row vector transpose of θ .

The symbol \parallel indicates the concatenation by columns of two matrices or vectors.

For example, if \mathbf{A} is a $p \times 1$ vector and \mathbf{B} is a $p \times 2$ matrix, then $\mathbf{C} = (\mathbf{A} \parallel \mathbf{B})$ is $p \times 3$.

The symbol // indicates the concatenation by rows of two matrices. For the above example, $\mathbf{C} = (\mathbf{A}^T \text{//} \mathbf{B}^T)^T$ is obtained as the transpose of the concatenation by rows of the two matrices \mathbf{A}^T and \mathbf{B}^T . Alternatively, $\mathbf{D} = (\mathbf{A} \text{//} \mathbf{A})$ is a $2p \times 1$ vector and $\mathbf{E} = (\mathbf{B} \text{//} \mathbf{B})$ is a $2p \times 2$ matrix.

The jk th element of the inverse of a matrix $[\mathbf{A}^{-1}]_{jk}$ is also expressed as \mathbf{A}^{jk} .

Let the $p \times p$ square matrix \mathbf{A} be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{rr} & \mathbf{A}_{rs} \\ \mathbf{A}_{sr} & \mathbf{A}_{ss} \end{bmatrix} \quad (\text{A.1})$$

with diagonal square matrices \mathbf{A}_{rr} and \mathbf{A}_{ss} of dimensions r and s , respectively, $r + s = p$. Then the partitioned inverse matrix is expressed as

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{rr} & \mathbf{A}^{rs} \\ \mathbf{A}^{sr} & \mathbf{A}^{ss} \end{bmatrix}, \quad (\text{A.2})$$

where, for example, \mathbf{A}^{rr} refers to the elements of \mathbf{A}^{-1} in the positions corresponding to the $r \times r$ submatrix, or $[\mathbf{A}^{-1}]_{rr}$. Thus, $\mathbf{A}^{rr} \neq \mathbf{A}_{rr}^{-1}$. For such a matrix, the upper left submatrix of the inverse is obtained as

$$\mathbf{A}^{rr} = [\mathbf{A}_{rr} - \mathbf{A}_{rs} \mathbf{A}_{ss}^{-1} \mathbf{A}_{sr}]^{-1}. \quad (\text{A.3})$$

A.1.3 Partition of Variation

A useful result is the principle of *partitioning of sums of squares* that forms the basis for the analysis of variance. Given a set of constants $\{a_i\}$ and values $\{z_i\}$ such that $\sum_i a_i y_i = \sum_i a_i z_i$, the total sum of squares of Y can be partitioned as

$$\sum_i a_i (y_i - \bar{y})^2 = \sum_i a_i (y_i - z_i)^2 + \sum_i a_i (z_i - \bar{y})^2 \quad (\text{A.4})$$

whenever it can be shown that $\sum_i a_i z_i (y_i - z_i) = 0$. The values z_i may be random observations or constants.

A similar result provides the *partitioning of variation (mean square error)* for an estimate $\hat{\theta}$ of a parameter θ . When the estimate has expectation $E(\hat{\theta}) \neq \theta$, then the mean square error of the estimate can be partitioned as

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 \\ &= V(\hat{\theta}) + [bias(\hat{\theta})]^2. \end{aligned} \quad (\text{A.5})$$

A related expression is the well-known result

$$V(Y) = E_X [V(Y|x)] + V_X [E(Y|x)], \quad (\text{A.6})$$

expressed in terms of the conditional moments of Y given the value of another variable X , integrated with respect to the distribution of X .

A.2 CENTRAL LIMIT THEOREM AND THE LAW OF LARGE NUMBERS

A.2.1 Univariate Case

The central limit theorem and the law of large numbers, when used in conjunction with other powerful theorems, such as Slutsky's theorem, provides the mechanism to derive the large sample or asymptotic distribution of many statistics, and virtually all of those employed in this book.

Let $\{y_1, y_2, \dots, y_N\}$ refer to a sample of N independent and identically distributed (*i.i.d.*) observations of the random variable Y drawn at random from a specified distribution with parameter vector θ , or

$$y_i \sim f(y; \theta) \quad \forall i \quad (\text{A.7})$$

for the density or probability mass distribution $f(y; \theta)$ with mean μ and variance σ^2 . Let \bar{y} be the sample mean, $\bar{y} = \sum_{i=1}^N y_i/N$. Then the *central limit theorem* is often expressed as

$$\bar{y} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/N) \quad (\text{A.8})$$

to indicate that the asymptotic or large sample distribution of the sample mean based on a large sample of size N is normal with expectation $E(\bar{y}) = \mu$ and variance $V(\bar{y}) = \sigma^2/N$. This is a casual notation that is fine in practice and thus is used throughout the text. The *large sample variance*, σ^2/N , of the sample mean \bar{y} may then be used as the basis for the computation of a large sample confidence interval or a statistical test. In the limit, however, this distribution is degenerate because $\lim_{n \rightarrow \infty} \sigma^2/n = 0$. Thus, a more statistically precise description is as follows:

Let $\{y_n\} = \{y_1, \dots, y_n\}$ denote a sequence of n *i.i.d.* observations drawn from a distribution $f(y; \theta)$ with mean μ and finite variance σ^2 . Let \bar{y}_n be the corresponding sequence of sample mean values computed from the first n observations. The *weak law of large numbers* specifies that the limiting value of the sample mean is the expected value of the random variable, often designated as the expectation of the i th observation in a sample; or that as $n \rightarrow \infty$, then

$$\bar{y}_n \xrightarrow{p} E(Y) = E(y_i) = \mu. \quad (\text{A.9})$$

Thus, the sample mean provides a *consistent estimator* of (converges in probability to) the mean of the distribution from which the observations arose.

Now let $S_n = (y_1 + \dots + y_n)$ be a sequence of partial sums of n *i.i.d.* observations drawn from this distribution, where $\bar{y}_n = S_n/n$. The *central limit theorem* then asserts that a sequence of partial sums computed as $n \rightarrow \infty$ satisfies

$$\lim_{n \rightarrow \infty} P \left[\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq z \right] = \Phi(z), \quad (\text{A.10})$$

where $\Phi(z)$ is the standard normal cumulative distribution function. Thus,

$$\lim_{n \rightarrow \infty} P \left[\frac{\bar{y}_n - \mu}{\sigma/\sqrt{n}} \leq z \right] = \Phi(z), \quad (\text{A.11})$$

where \bar{y}_n refers to the corresponding sequence of sample means as $n \rightarrow \infty$. Then, from (A.10),

$$\lim_{n \rightarrow \infty} P \left[\frac{1}{\sqrt{n}} \left(\frac{S_n - n\mu}{\sigma} \right) \leq z \right] = \Phi(z), \quad (\text{A.12})$$

and from (A.11),

$$\lim_{n \rightarrow \infty} P \left[\sqrt{n} \left(\frac{\bar{y}_n - \mu}{\sigma} \right) \leq z \right] = \Phi(z). \quad (\text{A.13})$$

Using $\mathcal{N}(.,.)$ to denote the normal distribution, the above expressions in terms of S_n and \bar{y}_n , respectively, can be expressed as

$$\frac{1}{\sqrt{n}} (S_n - n\mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad (\text{A.14})$$

and

$$\sqrt{n} (\bar{y}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad (\text{A.15})$$

to indicate convergence in distribution or in law to the normal distribution. These lead to the equivalent expressions

$$\frac{\bar{y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1), \quad (\text{A.16})$$

or as

$$\frac{(\bar{y}_n - \mu)^2}{\sigma^2/n} \xrightarrow{d} \chi^2_{(1)}, \quad (\text{A.17})$$

where $\chi^2_{(1)}$ refers to the central chi-square distribution on 1 *df*.

The expression in (A.15) implies that $P[(\bar{y}_n - \mu) > \epsilon] \xrightarrow{p} 0$ for any value ϵ which indicates that \bar{y} is a *consistent* estimator of the population mean μ regardless of the distribution $f(y; .)$ from which the observations are assumed drawn at random. A formal proof of the above results, generally known as the *Lindberg-Levy central limit theorem*, requires regularity conditions that essentially assert that the sequence of random variables $\{y_n\}$ has bounded variation, which is virtually always the case for the statistics considered herein.

The *Liapunov central limit theorem* (cf. Rao, 1973, p. 127) provides a generalization to the case of a sample of independent, though not identically distributed, random variables.

Expression (A.15) also demonstrates that \bar{y}_n is a \sqrt{n} -consistent estimator of μ . Any estimator $\hat{\theta}$ of a parameter θ is said to be \sqrt{n} -consistent if $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a nondegenerate probability distribution (cf. Lehmann, 1998). Such estimators converge to the true value at the rate of $n^{-\frac{1}{2}}$.

Example A.1 Simple Proportion

Consider the simple proportion of positive (+) responses from a sample of N *i.i.d.* Bernoulli observations where $Y = 1$ if +, 0 if -, with $E(Y) = P(y = 1) = \pi$ and $V(Y) = E(Y - \pi)^2 = \sigma^2 = \pi(1 - \pi)$. Then the sample proportion can be expressed as the mean of the $\{y_i\}$, or $p_n = \bar{y}_n = \sum_{i=1}^n y_i/n$ as $n \rightarrow \infty$, where $\sum_{i=1}^n y_i$ equals the number of + responses among the first n observations. Thus, the law of large numbers provides that the limiting value of the proportion is the probability, or that $p_n \xrightarrow{p} E(Y) = \pi$, and thus p is a consistent estimator for π . The central limit theorem further provides that

$$\sqrt{n} (p_n - \pi) \xrightarrow{d} \mathcal{N}[0, \pi(1 - \pi)], \quad (\text{A.18})$$

and asymptotically the large sample distribution of p with total sample size N is

$$p \approx \mathcal{N}[\pi, \pi(1 - \pi)/N]. \quad (\text{A.19})$$

A.2.2 Multivariate Case

Now consider a vector *i.i.d.* random variable $\mathbf{Y} = (Y_1 \dots Y_p)^T$ where the observation for the i th subject is a p -vector of measurements $\mathbf{y}_i = (y_{i1} \ y_{i2} \ \dots \ y_{ip})^T$, for $i = 1, \dots, n$, drawn from a p -variate distribution with mean vector $\boldsymbol{\mu} = (\mu_1 \ \dots \ \mu_p)^T$ and covariance matrix $\boldsymbol{\Sigma}$ of rank p . Let $\bar{\mathbf{y}}_n$ refer to a sequence of sample means computed as $n \rightarrow \infty$. Then, analogously to (A.9), the weak law of large numbers specifies that the limiting value of the sample mean vector is the mean vector

$$\bar{\mathbf{y}}_n \xrightarrow{p} \boldsymbol{\mu}, \quad (\text{A.20})$$

and analogously to (A.13), the central limit theorem specifies that

$$\lim_{n \rightarrow \infty} P \left[\sqrt{n} \boldsymbol{\Sigma}^{-1/2} (\bar{\mathbf{y}}_n - \boldsymbol{\mu}) \leq z \right] = \Phi_p(z), \quad (\text{A.21})$$

where $\Phi_p(z)$ refers to the standard p -variate normal cumulative distribution function. Using $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the p -variate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, this can be expressed as

$$\sqrt{n} \boldsymbol{\Sigma}^{-1/2} (\bar{\mathbf{y}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \quad (\text{A.22})$$

where \mathbf{I}_p is the identity matrix of order p . Thus,

$$\sqrt{n} (\bar{\mathbf{y}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad (\text{A.23})$$

and for a large sample of size N , the large sample distribution of the mean vector is asymptotically (approximately) distributed as

$$\bar{\mathbf{y}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/N). \quad (\text{A.24})$$

Further, from (A.23) the quadratic form

$$n(\bar{\mathbf{y}}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}}_n - \boldsymbol{\mu}) \xrightarrow{d} \chi_{(p)}^2, \quad (\text{A.25})$$

where $\chi_{(p)}^2$ is the central chi-square distribution on p df.

Example A.2 Multinomial Distribution

Consider the trinomial distribution of the frequencies $\mathbf{Y} = (Y_1 \ Y_2 \ Y_3)^T$ within each of three mutually exclusive categories from a sample of N *i.i.d.* observations, where Y_j is distributed as Bernoulli with probability π_j , $j = 1, 2, 3$. Then the i th observation is a vector $\mathbf{y}_i = (y_{i1} \ y_{i2} \ y_{i3})$, where y_{ij} is a binary variable that denotes whether the i th observation falls in the j th category; $y_{ij} = 1$ if j th category, 0 otherwise. Thus, $E(y_{ij}) = E(Y_j) = P(Y_j = 1) = \pi_j$ and $V(Y_j) = \sigma_j^2 = \pi_j(1 - \pi_j)$. It also follows that for any two categories $j \neq k$, $Cov(Y_j, Y_k) = \sigma_{jk} = E(Y_j Y_k) - E(Y_j)E(Y_k) = -\pi_j \pi_k$, since by construction, $y_{ij} y_{ik} = 0$. Thus, the \mathbf{y}_i have mean vector $\boldsymbol{\mu} = \boldsymbol{\pi} = (\pi_1 \ \pi_2 \ \pi_3)^T$ and covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \begin{bmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 \\ -\pi_1\pi_2 & \pi_2(1 - \pi_2) & -\pi_2\pi_3 \\ -\pi_1\pi_3 & -\pi_2\pi_3 & \pi_3(1 - \pi_3) \end{bmatrix}. \quad (\text{A.26})$$

The covariance matrix is often denoted as $\Sigma(\boldsymbol{\pi})$ because it depends explicitly on the mean vector $\boldsymbol{\pi}$. Since the \mathbf{Y} are subject to the linear constraint that $\sum_{j=1}^3 Y_j = 1$, the covariance matrix is singular with rank 2.

The vector of sample proportions can be expressed as the mean vector of the $\{\mathbf{y}_i\}$, or as $\mathbf{p} = (p_1 \ p_2 \ p_3)^T = \bar{\mathbf{y}} = \sum_{i=1}^n \mathbf{y}_i/N$, where $\sum_{i=1}^n \mathbf{y}_i$ is now the vector of frequencies within the three categories. Thus, from (A.24), for a large sample of size N , \mathbf{p} is asymptotically normally distributed with mean vector $\boldsymbol{\pi}$ and covariance matrix $\Sigma(\boldsymbol{\pi})/N$. Because these proportions are likewise subject to the linear constraint $\sum_{j=1}^3 p_j = 1$, the asymptotic distribution is degenerate with covariance matrix of rank 2. However, this poses no problems in practice because we need only characterize the distribution of any two of the three proportions.

A.3 DELTA METHOD

A.3.1 Univariate Case

A common problem in statistics is to derive the large sample moments, principally the expectation and variance, of a transformation of a statistic, including nonlinear transformations. These expressions are readily obtained by use of the *δ-method*.

Let T be any statistic for which the first two central moments are known, $E(T) = \mu$ and $V(T) = \sigma^2$. We desire the moments of a transformation $Y = g(T)$ for some function $g(\cdot)$ which is assumed to be twice differentiable, with derivatives designated as $g'(\cdot)$ and $g''(\cdot)$. A first-order *Taylor's series expansion* of $g(t)$ about μ is

$$g(t) = g(\mu) + g'(\mu)(t - \mu) + R_2(a). \quad (\text{A.27})$$

From the mean value theorem, the remainder to second order is

$$R_2(a) = (t - \mu)^2 g''(a)/2 \quad (\text{A.28})$$

for some value a contained in the interval (t, μ) . If the remainder vanishes under specified conditions, such as asymptotically, then

$$g(t) \cong g(\mu) + g'(\mu)(t - \mu), \quad (\text{A.29})$$

so that

$$E(Y) = \mu_y = E[g(t)] \cong g(\mu) + g'(\mu)E(t - \mu) = g(\mu) \quad (\text{A.30})$$

and

$$\begin{aligned} V(Y) &= E(Y - \mu_y)^2 \cong E[g(t) - g(\mu)]^2 \\ &= E[g'(\mu)(t - \mu)]^2 = [g'(\mu)]^2 V(T). \end{aligned} \quad (\text{A.31})$$

Herein, we frequently consider the moments of a transformation of a statistic T that is a consistent estimator of μ . In such cases, since $t \xrightarrow{P} \mu$, then the remainder in (A.28) vanishes asymptotically, or $R_2(a) \xrightarrow{P} 0$, and the above results apply to

any transformation of T . Furthermore, if $\widehat{V}(t)$ is a consistent estimator of $V(T)$, then it follows from Slutsky's convergence theorem (A.45, below), that $\widehat{V}(Y) = [g'(t)]^2 \widehat{V}(t)$ is a consistent estimator of $V(Y)$.

Example A.3 $\log(p)$

For example, consider the moments of the natural log of the simple proportion p for which $\mu = \pi$ and $\sigma^2 = \pi(1 - \pi)/N$. The Taylor's expansion yields

$$\log(p) = \log(\pi) + \frac{d \log(\pi)}{d\pi}(p - \pi) + R_2(a), \quad (\text{A.32})$$

where the remainder for some value $a \in (p, \pi)$ is $R_2(a) = (p - \pi)^2 g''(a)/2$. Since p is consistent for π , $p \xrightarrow{p} \pi$, then asymptotically $R_2(a) \rightarrow 0$ and thus

$$E[\log(p)] \cong \log(\pi), \quad (\text{A.33})$$

and

$$V[\log(p)] \cong \left[\frac{d \log(\pi)}{d\pi} \right]^2 V(p) = \left[\frac{1}{\pi^2} \right] \frac{\pi(1 - \pi)}{N} = \frac{1 - \pi}{\pi N}. \quad (\text{A.34})$$

A.3.2 Multivariate Case

Now consider a transformation of a p -vector $\mathbf{T} = (T_1 \cdots T_p)^T$ of statistics with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_T$. Assume that $\mathbf{Y} = (Y_1 \cdots Y_m)^T = \mathbf{G}(\mathbf{T}) = [g_1(\mathbf{T}) \cdots g_m(\mathbf{T})]^T$, $m \leq p$, where the k th transformation $g_k(\mathbf{T})$ is a twice-differentiable function of \mathbf{T} . Applying a first-order Taylor's series, as in (A.27), and assuming that the vector of remainders $\mathbf{R}_2(\mathbf{a})$ vanishes for values $\mathbf{a} \in (t, \boldsymbol{\mu})$, yields

$$E(\mathbf{Y}) = \boldsymbol{\mu}_Y \cong \mathbf{G}(\boldsymbol{\mu}) \quad (\text{A.35})$$

$$V(\mathbf{Y}) = \boldsymbol{\Sigma}_Y \cong \mathbf{H}(\boldsymbol{\mu})' \boldsymbol{\Sigma}_T \mathbf{H}(\boldsymbol{\mu}),$$

where $\mathbf{H}(\boldsymbol{\mu})$ is a $p \times m$ matrix with elements

$$\mathbf{H}(\boldsymbol{\mu}) = \left[\begin{array}{c} \partial g_1(\mathbf{t}) / \partial t \\ \vdots \\ \partial g_m(\mathbf{t}) / \partial t \end{array} \right]_{\mathbf{t}=\boldsymbol{\mu}}^T = \left[\begin{array}{ccc} \partial g_1(\mathbf{t}) / \partial t_1 & \cdots & \partial g_1(\mathbf{t}) / \partial t_p \\ \vdots & & \vdots \\ \partial g_m(\mathbf{t}) / \partial t_1 & \cdots & \partial g_m(\mathbf{t}) / \partial t_p \end{array} \right]_{\mathbf{t}=\boldsymbol{\mu}}^T \quad (\text{A.36})$$

evaluated at $\mathbf{t} = \boldsymbol{\mu}$.

When \mathbf{T} is a jointly consistent estimator for $\boldsymbol{\mu}$, then (A.35) provides the first two moments of the asymptotic distribution of \mathbf{Y} . Further, from Slutsky's theorem (A.45, below) if $\widehat{\boldsymbol{\Sigma}}_T$ is consistent for $\boldsymbol{\Sigma}_T$, then

$$\widehat{\boldsymbol{\Sigma}}_Y = \widehat{\mathbf{H}}(\mathbf{t})' \widehat{\boldsymbol{\Sigma}}_T \widehat{\mathbf{H}}(\mathbf{t}) = \widehat{\mathbf{H}}' \widehat{\boldsymbol{\Sigma}}_T \widehat{\mathbf{H}} \quad (\text{A.37})$$

is a consistent estimator of $\boldsymbol{\Sigma}_Y$.

Example A.4 Multinomial Generalized Logits

For example, consider the case of a trinomial where we wish to estimate the mean and variance of the vector of log odds (*logits*) of the second category versus the first, $\log(p_2/p_1)$, and also the third category versus the first, $\log(p_3/p_1)$. Thus, $\mathbf{p} = (p_1 \ p_2 \ p_3)^T$ has mean vector $\boldsymbol{\pi} = (\pi_1 \ \pi_2 \ \pi_3)^T$ and covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \frac{1}{N} \begin{bmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 \\ -\pi_1\pi_2 & \pi_2(1 - \pi_2) & -\pi_2\pi_3 \\ -\pi_1\pi_3 & -\pi_2\pi_3 & \pi_3(1 - \pi_3) \end{bmatrix}. \quad (\text{A.38})$$

The transformation is $\mathbf{Y} = \mathbf{G}(\mathbf{p}) = [g_1(\mathbf{p}) \ g_2(\mathbf{p})]^T$, where $g_1(\mathbf{p}) = \log(p_2/p_1)$ and $g_2(\mathbf{p}) = \log(p_3/p_1)$. Asymptotically, from (A.35),

$$E(\mathbf{Y}) = \boldsymbol{\mu}_Y = [\log(\pi_2/\pi_1) \ \log(\pi_3/\pi_1)]^T. \quad (\text{A.39})$$

To obtain the asymptotic variance requires the matrix of derivatives, which are

$$\begin{aligned} \mathbf{H}(\boldsymbol{\pi}) &= \begin{bmatrix} \partial g_1(\boldsymbol{\pi})/\partial\pi_1 & \partial g_1(\boldsymbol{\pi})/\partial\pi_2 & \partial g_1(\boldsymbol{\pi})/\partial\pi_3 \\ \partial g_2(\boldsymbol{\pi})/\partial\pi_1 & \partial g_2(\boldsymbol{\pi})/\partial\pi_2 & \partial g_2(\boldsymbol{\pi})/\partial\pi_3 \end{bmatrix}^T \\ &= \begin{bmatrix} -1/\pi_1 & 1/\pi_2 & 0 \\ -1/\pi_1 & 0 & 1/\pi_3 \end{bmatrix}^T. \end{aligned} \quad (\text{A.40})$$

Thus,

$$\boldsymbol{\Sigma}_Y = \mathbf{H}'\boldsymbol{\Sigma}(\boldsymbol{\pi})\mathbf{H} = \frac{1}{N} \begin{bmatrix} 1/\pi_1 + 1/\pi_2 & 1/\pi_1 \\ 1/\pi_1 & 1/\pi_1 + 1/\pi_3 \end{bmatrix} \quad (\text{A.41})$$

provides the asymptotic covariance matrix of the two logits.

A.4 SLUTSKY'S CONVERGENCE THEOREM

Slutsky's theorem (cf. Cramér, 1946; Serfling, 1980) is a multifaceted result which can be used to establish the convergence in distribution and/or the convergence in probability (consistency) of multidimensional transformations of a vector of statistics. For the purposes herein, I shall present these as two results rather than as a single theorem. The theorem is then used in conjunction with the delta method to obtain the asymptotic distribution of transformations of statistics.

A.4.1 Convergence in Distribution

The most common application of Slutsky's theorem concerns the asymptotic distribution of a linear combination of two sequences of statistics, one that converges in probability to a constant and another that converges in distribution to a specified distribution. In this book we are only concerned with functions of statistics that are asymptotically normally distributed, for which the theorem is so described.

The result, however, applies more generally to statistics that follow any specified distribution. The theorem also readily generalizes to more than two such statistics.

Let t_n be a sequence of statistics such that as $n \rightarrow \infty$,

$$\sqrt{n}(t_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad (\text{A.42})$$

where the variance σ^2 may be a function of the expectation μ . Also, let r_n be a sequence of statistics that converges in probability to a constant ρ , expressed as $r_n \xrightarrow{p} \rho$. Then

$$\sqrt{n}[(r_n + t_n) - (\rho + \mu)] \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad (\text{A.43})$$

and

$$\sqrt{n}(r_n t_n - \rho \mu) \xrightarrow{d} \mathcal{N}(0, \rho^2 \sigma^2). \quad (\text{A.44})$$

An example is provided below.

A.4.2 Convergence in Probability

Consider a set of statistics, each of which in sequence converges in probability to known quantities, such as $a_n \xrightarrow{p} \alpha$, $b_n \xrightarrow{p} \beta$, and $c_n \xrightarrow{p} \gamma$. For any continuous function $R(\cdot)$, then

$$R(a_n, b_n, c_n) \xrightarrow{p} R(\alpha, \beta, \gamma). \quad (\text{A.45})$$

This result provides the limiting expectation of a function of random variables and will be used to demonstrate the consistency of various "plug-in" estimators of parameters.

A.4.3 Convergence in Distribution of Transformations

The δ -method and Slutsky's theorem together provide powerful tools that can be used to derive the asymptotic distribution of statistics which are obtained as transformations of basic statistics known to be asymptotically normally distributed, such as transformations of means or proportions. The δ -method provides the expressions for the mean and variance of the statistic under a linear or nonlinear transformation. Slutsky's theorem then provides the proof of the asymptotic convergence to a normal, or multivariate normal, distribution. This result is stated herein for the univariate case; the multivariate case follows similarly (cf. Rao, 1973, p. 385; or Bickel and Doksum, 1977, pp. 461–462). Under certain conditions, similar results can be used to demonstrate convergence to other distributions. However, herein we only consider the application to transformations of statistics that converge in distribution to a normal distribution.

As above, let $\sqrt{n}(t_n - \mu)$ be a sequence of statistics which converges in distribution to the normal distribution with mean 0 and variance σ^2 as in (A.42). Also, let $g(t)$ be a single variable function with derivative $g'(\cdot)$ that is continuous at μ . Then $g(t)$ converges in distribution to

$$\sqrt{n}[g(t_n) - g(\mu)] \xrightarrow{d} \mathcal{N}\left(0, [g'(\mu)]^2 \sigma^2\right). \quad (\text{A.46})$$

Thus, for large N , the approximate large sample distribution of $g(t)$ is

$$g(t) \xrightarrow{d} \mathcal{N} \left(g(\mu), \frac{[g'(\mu)]^2 \sigma^2}{N} \right). \quad (\text{A.47})$$

Example A.5 $\log(p)$

Consider the asymptotic distribution of $\log(p)$. Applying the Taylor's expansion (A.27), then from (A.32) asymptotically

$$\sqrt{n} [\log(p) - \log(\pi)] = \sqrt{n} \frac{(p - \pi)}{\pi} - \sqrt{n} \frac{(p - \pi)^2}{2a^2}, \quad (\text{A.48})$$

where the last term is $\sqrt{n}R_2(a)$. Since p is asymptotically normally distributed, then the first term on the right hand side is likewise asymptotically normally distributed.

To evaluate the remainder asymptotically, let $\{p_n\}$ denote a sequence of values as $n \rightarrow \infty$. In Example A.1 we saw that p_n is a sample mean of n Bernoulli variables, and thus is a \sqrt{n} -consistent estimator of π so that $(p_n - \pi)^2 \rightarrow 0$ faster than $n^{-\frac{1}{2}}$, and thus $\sqrt{n}R_2(a) \xrightarrow{p} 0$. Therefore, asymptotically $\sqrt{n} [\log(p) - \log(\pi)]$ is the sum of two random variables, one converging in distribution to the normal, the other converging in probability to a constant (zero). From Slutsky's convergence in distribution theorem (A.43), it follows that

$$\sqrt{n} [\log(p) - \log(\pi)] \xrightarrow{d} \mathcal{N} \left[0, \left(\frac{d \log(\pi)}{d\pi} \right)^2 \pi(1 - \pi) \right] \quad (\text{A.49})$$

and for large N asymptotically,

$$\log(p) \xrightarrow{d} \mathcal{N} \left(\log(\pi), \frac{1 - \pi}{N\pi} \right). \quad (\text{A.50})$$

The large sample variance of $\log(p)$ is

$$V [\log(p)] \cong \frac{1 - \pi}{N\pi}, \quad (\text{A.51})$$

which can be estimated by "plugging in" the estimate p for π to obtain

$$\widehat{V} [\log(p)] = \frac{1 - p}{Np}. \quad (\text{A.52})$$

Then from Slutsky's convergence in probability theorem (A.45), since $p_n \xrightarrow{p} \pi$ it follows that (A.52) \xrightarrow{p} (A.51), which proves the consistency of the large sample estimate of $V [\log(p)]$.

From the asymptotic distribution of $\log(p)$ in (A.50) and the consistency of the estimate of the variance using the "plug-in" approach in (A.52), again using Slutsky's convergence in distribution theorem (A.44), it follows that asymptotically

$$\frac{\log(p) - \log(\pi)}{\sqrt{\widehat{V} [\log(p)]}} \xrightarrow{d} N(0, 1). \quad (\text{A.53})$$

Thus, the asymptotic coverage probability of $(1 - \alpha)$ level confidence limits based on the estimated large sample variance in (A.52) is approximately the desired level $1 - \alpha$.

Example A.6 Multinomial Logits

For the multinomial logits in Example A.4, it follows from the above theorems that the sample logits $\mathbf{Y} = [\log(p_2/p_1) \log(p_3/p_1)]^T$ are distributed as bivariate normal with expectation vector $\mu_Y = [\log(\pi_2/\pi_1) \log(\pi_3/\pi_1)]^T$ and with variance as presented in (A.41). Substituting the elements $\mathbf{p} = (p_1 \ p_2 \ p_3)^T$ for $\boldsymbol{\pi} = (\pi_1 \ \pi_2 \ \pi_3)^T$ yields the estimated large sample covariance matrix

$$\widehat{\Sigma}_Y = \frac{1}{n} \begin{bmatrix} 1/p_1 + 1/p_2 & 1/p_1 \\ 1/p_1 & 1/p_1 + 1/p_3 \end{bmatrix}. \quad (\text{A.54})$$

Since \mathbf{p} is jointly consistent for $\boldsymbol{\pi}$, then from Slutsky's convergence in probability theorem (A.45), $\widehat{\Sigma}_Y$ is a consistent estimate of Σ_Y . This estimate may also be obtained by evaluating

$$\widehat{\Sigma}_Y = \widehat{\mathbf{H}}' \widehat{\Sigma}(\mathbf{p}) \widehat{\mathbf{H}}, \quad (\text{A.55})$$

in which $\widehat{\Sigma}(\mathbf{p}) = \Sigma(\boldsymbol{\pi})_{|\boldsymbol{\pi}=\mathbf{p}}$ and in which $\widehat{\mathbf{H}} = \mathbf{H}(\mathbf{p}) = \mathbf{H}(\boldsymbol{\pi})_{|\boldsymbol{\pi}=\mathbf{p}}$ as in (A.37).

A.5 LEAST SQUARES ESTIMATION

A.5.1 Ordinary Least Squares

Ordinary least squares (OLS) is best known as the method for the derivation of the parameter estimates in the simple linear regression model. In general, OLS can be described as a method for the estimation of the conditional expectation of the dependent variable Y within the context of some model as a function of a covariate vector with value \mathbf{x} , that may represent a vector of p covariates, plus a constant (1) for the intercept. The values of \mathbf{x} are considered fixed quantities; that is, we condition on the observed values of \mathbf{x} . In the simple linear multiple regression model, the conditional expectation is expressed as $E(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\theta}$ as a function of a vector of parameters $\boldsymbol{\theta}$. This specification is termed the *structural component* of the regression model, which in this instance is a linear function of the covariates. Given \mathbf{x} , it is then assumed that the random errors $\varepsilon = y - E(Y|\mathbf{x})$ are statistically independent and identically distributed (*i.i.d.*) with mean zero and common variance σ_ε^2 . This specification is termed the *random component* of the regression model. The structural and random components together specify that $y = E(Y|\mathbf{x}) + \varepsilon$, which then specifies that $V(Y|\mathbf{x}) = \sigma_\varepsilon^2$ conditionally on \mathbf{x} .

For a sample of N observations, let $\mathbf{Y} = (y_1 \ \cdots \ y_N)^T$ refer to the column vector of dependent variable values, and \mathbf{X} refer to the $N \times (p+1)$ vector of covariate values for the N observations. The i th row of \mathbf{X} consists of the vector of $(p+1)$ covariate values for the i th observation, $\mathbf{x}_i^T = (1 \ x_{i1} \ \cdots \ x_{ip})$. For the i th observation, the random error is $\varepsilon_i = y_i - E(Y|\mathbf{x}_i)$, and for the sample of N observations the

vector of random errors is $\boldsymbol{\epsilon} = (\epsilon_1 \cdots \epsilon_N)^T$, where, as stated above, it is assumed that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $Cov(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma_{\boldsymbol{\epsilon}}^2 \mathbf{I}_N$, with \mathbf{I}_N being the identity matrix of dimension N . In vector notation, the linear model then can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (\text{A.56})$$

where $\boldsymbol{\theta} = (\alpha // \boldsymbol{\beta})$, $\boldsymbol{\beta} = (\beta_1 \cdots \beta_p)^T$, α being the intercept. Note that throughout the Appendix the intercept is implicit in the expression $\mathbf{x}'\boldsymbol{\theta}$, whereas elsewhere in the book we use the explicit notation $\alpha + \mathbf{x}'\boldsymbol{\beta}$.

We then desire an estimate of the coefficient vector $\boldsymbol{\theta}$ that satisfies some desirable statistical properties. In this setting squared error loss suggests choosing the vector $\boldsymbol{\theta}$ so as to minimize the sums of squares of errors $\sum_j \epsilon_j^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = SSE$. Thus, the estimate satisfies

$$\min_{\hat{\boldsymbol{\theta}}} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = \min_{\hat{\boldsymbol{\theta}}} (\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}}). \quad (\text{A.57})$$

Using the calculus of maxima/minima, the solution is obtained by setting the vector of first derivatives equal to $\mathbf{0}$ and solving for $\hat{\boldsymbol{\theta}}$. For this purpose we require the derivatives of the bilinear and quadratic forms with respect to $\hat{\boldsymbol{\theta}}$ as follows:

$$\frac{\partial \mathbf{X}\hat{\boldsymbol{\theta}}}{\partial \hat{\boldsymbol{\theta}}} = \mathbf{X}, \quad \frac{\partial \hat{\boldsymbol{\theta}}'\mathbf{A}\hat{\boldsymbol{\theta}}}{\partial \hat{\boldsymbol{\theta}}} = 2\hat{\boldsymbol{\theta}}'\mathbf{A}. \quad (\text{A.58})$$

Thus,

$$\frac{\partial SSE}{\partial \hat{\boldsymbol{\theta}}} = \frac{\partial (\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} = -2\mathbf{Y}'\mathbf{X} + 2\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}, \quad (\text{A.59})$$

and the matrix of second derivatives ($2\mathbf{X}'\mathbf{X}$) is positive definite, provided that it is of full rank. Thus, setting the vector of first derivatives equal to $\mathbf{0}$ yields the OLS estimating equation

$$\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X} - \mathbf{Y}'\mathbf{X} = \mathbf{0}, \quad (\text{A.60})$$

for which the SSE is minimized with respect to the choice of $\hat{\boldsymbol{\theta}}$. Solving for $\hat{\boldsymbol{\theta}}$ yields the OLS estimate

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}. \quad (\text{A.61})$$

A unique solution vector for $\hat{\boldsymbol{\theta}}$ is obtained provided that $\mathbf{X}'\mathbf{X}$ is of full rank and the inverse exists. Another basic result from the algebra of matrices is that the rank of $\mathbf{X}'\mathbf{X}$ is the minimum of the row and column rank of \mathbf{X} . Since \mathbf{X} is a $n \times (p+1)$ matrix, then $\mathbf{X}'\mathbf{X}$ will be positive definite of full rank $(p+1)$ unless there is a linear dependency or degeneracy among the columns of the \mathbf{X} matrix, which would require that one covariate in the design matrix \mathbf{X} be a linear combination of the others. If such a degeneracy exists, then there is no unique solution vector $\hat{\boldsymbol{\theta}}$ that satisfies the OLS estimating equation. In this case, one among the many solutions is obtained by using a generalized inverse. Throughout, however, we assume that $\mathbf{X}'\mathbf{X}$ is of full rank, unless stated otherwise.

A.5.2 Gauss-Markov Theorem

The properties of least squares estimators are provided by the *Gauss-Markov theorem*. From the assumption of *i.i.d.* homoscedastic errors (with common variance), the following properties are readily derived (cf. Rao, 1973).

The least squares estimates of the coefficients are unbiased, since

$$E(\hat{\boldsymbol{\theta}}) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' E(\mathbf{Y}) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} \boldsymbol{\theta} = \boldsymbol{\theta}. \quad (\text{A.62})$$

To obtain the variance of the estimates, note that the solution $\hat{\boldsymbol{\theta}}$ is a linear combination of the \mathbf{Y} vector of the form $\hat{\boldsymbol{\theta}} = \mathbf{H}' \mathbf{Y}$, where $\mathbf{H}' = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$. Since $V(\mathbf{Y}) = V(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_N$, then

$$\begin{aligned} V(\hat{\boldsymbol{\theta}}) &= \mathbf{H}' [V(\mathbf{Y})] \mathbf{H} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' [\sigma_\varepsilon^2 \mathbf{I}_N] \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \sigma_\varepsilon^2. \end{aligned} \quad (\text{A.63})$$

Since

$$\hat{\sigma}_\varepsilon^2 = MSE = \frac{(\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\theta}})' (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\theta}})}{N - p - 1} \quad (\text{A.64})$$

provides a consistent estimator of σ_ε^2 , then a consistent estimator of the covariance matrix of the estimates is provided by

$$\hat{V}(\hat{\boldsymbol{\theta}}) = (\mathbf{X}' \mathbf{X})^{-1} \hat{\sigma}_\varepsilon^2. \quad (\text{A.65})$$

Since the least squares estimator is that linear function of the observations which is both unbiased and which minimizes the *SSE*, or which has the smallest variance among all possible unbiased linear estimators, then this is a *best linear unbiased estimator (BLUE)*. Again, this result only requires the assumption that the $\{y_i\}$ are independent and that the random errors are *i.i.d.* with mean zero and constant variance σ_ε^2 conditional on the covariates $\{\mathbf{x}_i\}$. No further assumptions are necessary.

In addition, if it is assumed that the $\{\varepsilon_i\}$ are normally distributed, then the *F* distribution can be used to characterize the distribution of a ratio of independent sums of squares as the basis for parametric tests with finite samples. However, the normal errors assumption is not necessary for a large sample inference. Since the $\{y_i\}$ are independent with constant conditional variance σ_ε^2 given $\{\mathbf{x}_i\}$, then from (A.62) and the Liapunov central limit theorem, asymptotically

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N} \left[\mathbf{0}, V(\hat{\boldsymbol{\theta}}) \right]. \quad (\text{A.66})$$

This provides the basis for large sample Wald tests and confidence limits for the elements of $\boldsymbol{\theta}$.

A.5.3 Weighted Least Squares

Weighted least squares (WLS) generalizes these developments to the case where the random errors have expectation zero but are not *i.i.d.*, meaning that $V(\boldsymbol{\varepsilon}) =$

$\Sigma_\epsilon \neq \sigma_\epsilon^2 \mathbf{I}_N$. Thus, there may be heteroscedasticity of the error variances such that $V(\epsilon_i) \neq V(\epsilon_j)$ for some two observations $i \neq j$, and/or the errors may be correlated such that $Cov(\epsilon_i, \epsilon_j) \neq 0$. As with ordinary least squares, we start with the assumption of a simple linear multiple regression model such that the conditional expectation is expressed as $E(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\theta}$. Thus, the model specifies that $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ but where $V(\boldsymbol{\epsilon}) = \Sigma_\epsilon$. When Σ_ϵ is of full rank, then by a transformation using the root of the inverse, $\Sigma_\epsilon^{-1/2}$, we have

$$\Sigma_\epsilon^{-1/2}\mathbf{Y} = \Sigma_\epsilon^{-1/2}\mathbf{X}\boldsymbol{\theta} + \Sigma_\epsilon^{-1/2}\boldsymbol{\epsilon}, \quad (\text{A.67})$$

which can be expressed as

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\theta} + \tilde{\boldsymbol{\epsilon}}, \quad (\text{A.68})$$

where

$$V(\tilde{\boldsymbol{\epsilon}}) = \Sigma_\epsilon^{-1/2}\Sigma_\epsilon\Sigma_\epsilon^{-1/2} = \mathbf{I}_N. \quad (\text{A.69})$$

Thus, the transformed random errors $\tilde{\boldsymbol{\epsilon}}$ satisfy the assumptions of the OLS estimators so that

$$\hat{\boldsymbol{\theta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} = (\mathbf{X}'\Sigma_\epsilon^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Sigma_\epsilon^{-1}\mathbf{Y}). \quad (\text{A.70})$$

As in (A.62)–(A.63), it follows that the WLS estimates are unbiased with covariance matrix

$$V(\hat{\boldsymbol{\theta}}) = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} = (\mathbf{X}'\Sigma_\epsilon^{-1}\mathbf{X})^{-1}, \quad (\text{A.71})$$

and that the estimates are asymptotically normally distributed as in (A.66). Further, if a consistent estimator $\hat{\Sigma}_\epsilon$ of Σ_ϵ is available, then the covariance matrix of the estimates can be consistently estimated as

$$\hat{V}(\hat{\boldsymbol{\theta}}) = (\mathbf{X}'\hat{\Sigma}_\epsilon^{-1}\mathbf{X})^{-1}, \quad (\text{A.72})$$

which provides the basis for large sample confidence intervals and tests of significance. In these expressions, the matrix Σ_ϵ^{-1} is termed the *weight matrix* since the solution vector $\hat{\boldsymbol{\theta}}$ minimizes the weighted *SSE*, computed as

$$SSE = (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})'(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})'\Sigma_\epsilon^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}). \quad (\text{A.73})$$

Example A.7 Heteroscedasticity

One common application of weighted least squares is the case of heteroscedastic variances. One instance is where y_i is the mean of n_i measurements for the i th observation, where n_i varies among the observations. Thus,

$$\Sigma_\epsilon = \sigma_\epsilon^2[\text{diag}(n_1^{-1} \cdots n_N^{-1})]. \quad (\text{A.74})$$

Example A.8 Correlated Observations

Another common application is to the case of clustered sampling, such as repeated or clustered measures on the same subject, or measures on the members of a family. In this case the sampling unit is the cluster and the observations within a cluster are correlated. Let n_j refer to the number of members within the j th cluster and let Σ_j denote the $n_j \times n_j$ matrix of the variances and covariances among the n_j observations within the j th cluster. Then

$$\Sigma_{\epsilon} = \text{blockdiag}[\Sigma_1 \cdots \Sigma_N] = \begin{bmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & \Sigma_N \end{bmatrix}, \quad (\text{A.75})$$

where N is now the total number of clusters. The within-cluster variances and covariances may be specified on the basis of the statistical properties of the sampling procedure, or may be estimated from the data.

For example, for n_j repeated measures on the same subject, one may assume an *exchangeable correlation* structure such that there is constant correlation, say ρ , among measures within the subject (cluster). Given an estimate of the common correlation, $\hat{\rho}$, and estimates of the variances of each repeated measure, say $\hat{\sigma}_{\ell}^2$, then the covariance between the first and second repeated measures, for example, is $\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2$. From these the estimated covariance matrix could be obtained for the set of repeated measures for the j th subject.

A.5.4 Iteratively Reweighted Least Squares

A further generalization, iteratively reweighted least squares (IRLS), is employed in instances where the covariance matrix of the random errors is a function of the estimated conditional expectations, such that $\Sigma_{\epsilon} = \mathbf{G}(\boldsymbol{\theta})$ for some $N \times N$ matrix with elements that are a function of the coefficient vector $\boldsymbol{\theta}$. In this case, the weight matrix depends on the values of the coefficients. In general, an iterative procedure is required as follows.

Given some initial estimates of the coefficients, say $\hat{\boldsymbol{\theta}}_0$, one computes the weight matrix, say $\mathbf{G}_0 = \mathbf{G}(\hat{\boldsymbol{\theta}}_0)$, and then computes the first-step estimates as

$$\hat{\boldsymbol{\theta}}_1 = (\mathbf{X}' \mathbf{G}_0^{-1} \mathbf{X})^{-1} (\mathbf{X}' \mathbf{G}_0^{-1} \mathbf{Y}). \quad (\text{A.76})$$

Then the updated weight matrix $\mathbf{G}_1 = \mathbf{G}(\hat{\boldsymbol{\theta}}_1)$ and the second-step estimate $\hat{\boldsymbol{\theta}}_2$ are obtained. The iterative process continues until the coefficient estimates converge to a constant vector, or equivalently until some objective function such as the *SSE* converges to a constant. As with ordinary least squares, the final estimates are asymptotically normally distributed as in (A.66), where a consistent estimate of the covariance matrix of the coefficient estimates is provided by (A.72) using $\hat{\Sigma}_{\epsilon} = \mathbf{G}(\hat{\boldsymbol{\theta}})$.

Often a relationship between the conditional expectations and the variance of the errors arises from the specification of a parametric model in the population based on

an underlying distribution conditional on the values of the covariates. In this case the IRLS estimates of the coefficients equal those obtained from maximum likelihood estimation.

A.6 MAXIMUM LIKELIHOOD ESTIMATION AND EFFICIENT SCORES

A.6.1 Estimating Equation

Another statistical approach to the estimation of parameters and the development of statistical tests is to adopt a likelihood based on an assumed underlying population model. Let (y_1, y_2, \dots, y_N) refer to a sample of independent and identically distributed (*i.i.d.*) observations drawn at random from a specified distribution $f(y; \theta)$, where the density or probability mass distribution $f(\cdot)$ has unknown parameter θ , which may be a p -vector of parameters $\theta = (\theta_1 \dots \theta_p)^T$. Throughout we assume that θ is the true parameter value, although at times we use the notation $\theta = \theta_0$ to denote that the true value is assumed to be the specified value θ_0 . When not ambiguous, results are presented for the scalar case using the parameter θ .

The *likelihood function* then is the total probability of the sample under the assumed model

$$L(y_1, \dots, y_N; \theta) = \prod_{i=1}^N f(y_i; \theta), \quad (\text{A.77})$$

which, for simplicity, is designated as simply $L(\theta)$. Alternatively, when a known sufficient statistic, say T , is known to exist for θ , then apart from constants, the likelihood may be expressed in terms of the distribution of T .

The *maximum likelihood estimate* of θ , designated as $\hat{\theta}$, is that value for which the likelihood is maximized. This value is most easily determined using the log likelihood, which in the case of (A.77), is represented as a sum rather than a product of terms:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^N \log f(y_i; \theta). \quad (\text{A.78})$$

The *MLE* $\hat{\theta}$ is then obtained from the calculus of the local extrema (max/min) of a function. Thus, in the single parameter case, the maximum likelihood estimator (*MLE*) is that value such that

$$\frac{d\ell(\theta)}{d\theta} \Big|_{\theta=\hat{\theta}} = 0 \quad \text{given that} \quad \frac{d^2\ell(\theta)}{d\theta^2} \Big|_{\theta=\hat{\theta}} < 0. \quad (\text{A.79})$$

The first derivative is the slope of the function $\ell(\theta)$ with respect to θ . The maximum occurs at the point where the tangent to $\ell(\theta)$ is the horizontal line or the point along the likelihood surface where the slope is zero. The second derivative represents the degree of curvature of the function $\ell(\theta)$ and its direction: facing up or down. Thus, the condition on the second derivative is that the likelihood function be convex or "concave down."

The *maximum likelihood estimating equations*, therefore, in the scalar or vector parameter cases are

$$\frac{d\ell(\theta)}{d\theta} = 0 \quad \text{or} \quad \begin{pmatrix} \frac{\partial\ell(\theta)}{\partial\theta_1} \\ \vdots \\ \frac{\partial\ell(\theta)}{\partial\theta_p} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (\text{A.80})$$

The *MLE* $\hat{\theta}$ then is the solution for θ in the scalar parameter case, and $\hat{\theta}$ is the solution for θ in the multiparameter case, the vector estimating equation applying simultaneously to the p elements of θ . In many cases the solution of these estimating equations requires an iterative procedure such as the Newton-Raphson method.

A.6.2 Efficient Score

The estimating equation for a scalar θ is a function of the *Fisher efficient score*

$$U(\theta) = \frac{d\ell(\theta)}{d\theta} = \sum_{i=1}^N \frac{d \log f(y_i; \theta)}{d\theta} = \sum_{i=1}^N U_i(\theta), \quad (\text{A.81})$$

where $U_i(\theta)$ is the Fisher efficient score for the i th observation. When θ is a p -vector then the total score vector is

$$\mathbf{U}(\theta) = [U(\theta)_{\theta_1} \ \cdots \ U(\theta)_{\theta_p}]^T = \sum_{i=1}^N \mathbf{U}_i(\theta), \quad (\text{A.82})$$

where for the i th subject,

$$\begin{aligned} \mathbf{U}_i(\theta) &= [U_i(\theta)_{\theta_1} \ \cdots \ U_i(\theta)_{\theta_p}]^T \\ &= \left[\frac{\partial \log f(y_i; \theta)}{\partial \theta_1} \ \cdots \ \frac{\partial \log f(y_i; \theta)}{\partial \theta_p} \right]^T. \end{aligned} \quad (\text{A.83})$$

The notation $U(\theta)_{\theta_j}$ designates that the score for the j th element of the parameter vector is some function of the p -vector θ . The *MLE* is defined as that value of the p -vector $\hat{\theta}$ for which the total score is the zero vector.

An important property of the efficient score is that $E[\mathbf{U}(\theta)] = \mathbf{0}$ when the score is evaluated at the true value of the parameter in the population θ . To derive this result, consider the case where θ is a scalar. Since

$$U_i(\theta) = \frac{d \log f(y_i; \theta)}{d\theta} = \frac{1}{f(y_i; \theta)} \frac{df(y_i; \theta)}{d\theta}, \quad (\text{A.84})$$

then

$$\begin{aligned} E[U_i(\theta)] &= \int \frac{1}{f(y_i; \theta)} \frac{df(y_i; \theta)}{d\theta} f(y_i; \theta) dy_i = \int \frac{df(y_i; \theta)}{d\theta} dy_i \\ &= \frac{d}{d\theta} \int f(y_i; \theta) dy_i = \frac{d(1)}{d\theta} = 0 \quad \forall i. \end{aligned} \quad (\text{A.85})$$

Thus,

$$E[U(\theta)] = E \left[\sum_{i=1}^N U_i(\theta) \right] = 0. \quad (\text{A.86})$$

This result also applies to the multiparameter case where $E[U(\theta)] = 0$.

This property plays a central role in the development of the efficient score test of a hypothesis of the form $H_0: \theta = \theta_0$.

A.6.3 Fisher's Information Function

Again consider the case where θ is a scalar. Since the likelihood in (A.77) is the probability of any observed sample of N *i.i.d.* observations, then

$$\int \cdots \int L(y_1, \dots, y_N; \theta) dy_1 \cdots dy_N = 1. \quad (\text{A.87})$$

Taking the derivative with respect to θ yields

$$\int \cdots \int \frac{dL(\theta)}{d\theta} dy_1 \cdots dy_N = 0. \quad (\text{A.88})$$

As in (A.84), however,

$$\frac{d\ell(\theta)}{d\theta} = \frac{d \log L(\theta)}{d\theta} = \frac{1}{L(\theta)} \frac{dL(\theta)}{d\theta}, \quad (\text{A.89})$$

so that

$$\frac{dL(\theta)}{d\theta} = \frac{d\ell(\theta)}{d\theta} L(\theta). \quad (\text{A.90})$$

Then, (A.88) can be expressed as

$$\int \cdots \int \frac{d\ell(\theta)}{d\theta} L(\theta) dy_1 \cdots dy_N = E[U(\theta)] = 0, \quad (\text{A.91})$$

which is a generalization of (A.85).

Differentiating $E[U(\theta)]$ in (A.91) a second time with respect to θ yields

$$\begin{aligned} \frac{dE[U(\theta)]}{d\theta} &= \int \cdots \int \left[\frac{d\ell(\theta)}{d\theta} \left(\frac{dL(\theta)}{d\theta} \right) + L(\theta) \left(\frac{d^2\ell}{d\theta^2} \right) \right] dy_1 \cdots dy_N = 0 \\ &= \int \cdots \int \left[\left(\frac{d\ell}{d\theta} \right)^2 L(\theta) + \left(\frac{d^2\ell}{d\theta^2} \right) L(\theta) \right] dy_1 \cdots dy_N = 0, \end{aligned} \quad (\text{A.92})$$

so that

$$\frac{dE[U(\theta)]}{d\theta} = E \left[\left(\frac{d\ell}{d\theta} \right)^2 \right] + E \left[\frac{d^2\ell}{d\theta^2} \right] = 0. \quad (\text{A.93})$$

Since the two terms sum to zero and the first term must be positive, then this yields *Fisher's information equality*

$$\begin{aligned} I(\theta) &= E \left[\left(\frac{d\ell}{d\theta} \right)^2 \right] = E [U(\theta)^2] \\ &= E \left[-\frac{d^2\ell}{d\theta^2} \right] = E [-U'(\theta)]. \end{aligned} \quad (\text{A.94})$$

The *information function* $I(\theta)$ quantifies the expected amount of information in a sample of N observations concerning the true value of θ . The second derivative of the likelihood function, or of the log likelihood, with respect to θ describes the curvature of the likelihood in the neighborhood of θ . Thus, the greater the negative derivative, the sharper is the peak of the likelihood function, and the less dispersed the likelihood over the parameter space of θ . Thus, the greater is the information about the true value of θ .

In the general multiparameter case, *Fisher's information function* for a p -vector θ may be defined in terms of the matrix of mixed partial second derivatives, which is a $p \times p$ matrix defined as

$$\mathbf{I}(\theta) = E \left[-\frac{\partial^2\ell}{\partial\theta^2} \right] \quad \text{with elements} \quad \mathbf{I}(\theta)_{jk} = E \left[-\frac{\partial^2\ell}{\partial\theta_j\partial\theta_k} \right] \quad (\text{A.95})$$

for $1 \leq j \leq k \leq p$. Alternatively, $\mathbf{I}(\theta)$ may be defined from the outer product of the score vector as

$$\mathbf{I}(\theta) = E \left[\left(\frac{\partial\ell}{\partial\theta} \right) \left(\frac{\partial\ell}{\partial\theta} \right)^T \right] \quad \text{with elements} \quad \mathbf{I}(\theta)_{jk} = E \left[\left(\frac{\partial\ell}{\partial\theta_j} \right) \left(\frac{\partial\ell}{\partial\theta_k} \right) \right]. \quad (\text{A.96})$$

These expressions describe the *expected information function*.

The matrix of mixed second derivatives is commonly known as the *Hessian* matrix,

$$\mathbf{H}(\theta) = \frac{\partial^2\ell}{\partial\theta^2} = \left\{ \frac{\partial^2\ell}{\partial\theta_j\partial\theta_k} \right\} = \frac{\partial\mathbf{U}(\theta)}{\partial\theta} = \mathbf{U}'(\theta) = \sum_{i=1}^N \mathbf{U}'_i(\theta). \quad (\text{A.97})$$

Thus, from (A.95), $\mathbf{I}(\theta) = E[-\mathbf{H}(\theta)]$. The *observed information*, therefore, is

$$\mathbf{i}(\theta) = -\mathbf{H}(\theta) = -\mathbf{U}'(\theta). \quad (\text{A.98})$$

For a sample of *i.i.d.* observations,

$$\mathbf{I}(\theta) = E[\mathbf{i}(\theta)] = -E \left[\sum_{i=1}^N \mathbf{U}'_i(\theta) \right] = -NE[\mathbf{U}'_i(\theta)] \quad (\text{A.99})$$

for any randomly selected observation (the i th). Thus, if $E[\mathbf{U}'_i(\theta)]$ exists in closed form, in terms of θ , then one can derive the expression for the expected information $\mathbf{I}(\theta)$ for any true value θ .

The observed and expected information can also be expressed in terms of the sums of squares and cross products obtained from the outer product of the vectors of efficient scores. From (A.96) the $p \times p$ matrix of outer products is

$$\begin{aligned} \left(\frac{\partial \ell}{\partial \theta} \right) \left(\frac{\partial \ell}{\partial \theta} \right)^T &= \mathbf{U}(\theta) \mathbf{U}(\theta)^T = \left[\sum_{i=1}^N \mathbf{U}_i(\theta) \right] \left[\sum_{i=1}^N \mathbf{U}_i(\theta)^T \right] \quad (\text{A.100}) \\ &= \sum_{i=1}^N \mathbf{U}_i(\theta) \mathbf{U}_i(\theta)^T + \sum_{i \neq j} \mathbf{U}_i(\theta) \mathbf{U}_j(\theta)^T. \end{aligned}$$

Therefore,

$$\begin{aligned} E \left[\left(\frac{\partial \ell}{\partial \theta} \right) \left(\frac{\partial \ell}{\partial \theta} \right)^T \right] &= E \left[\sum_{i=1}^N \mathbf{U}_i(\theta) \mathbf{U}_i(\theta)^T \right] \quad (\text{A.101}) \\ &+ E \left[\sum_{i \neq j} \mathbf{U}_i(\theta) \mathbf{U}_j(\theta)^T \right]. \end{aligned}$$

Since the observations are *i.i.d.*, then

$$E [\mathbf{U}_i(\theta) \mathbf{U}_j(\theta)^T] = E [\mathbf{U}_i(\theta)] E [\mathbf{U}_j(\theta)^T] = \mathbf{0} \quad (\text{A.102})$$

for all $1 \leq i < j \leq N$. Therefore,

$$I(\theta) = E [\mathbf{i}(\theta)] = E \left[\sum_{i=1}^N \mathbf{U}_i(\theta) \mathbf{U}_i(\theta)^T \right] = N E [\mathbf{U}_i(\theta) \mathbf{U}_i(\theta)^T] \quad (\text{A.103})$$

for any random observation, arbitrarily designated as the *i*th.

A.6.4 Cramér-Rao Inequality: Efficient Estimators

Developments similar to the above lead to an important result which establishes the lower bound for the variance of an estimator, the Cramér-Rao lower bound. Consider a statistic T that provides an unbiased estimate of some function $\mu(\theta)$ of the scalar parameter θ such that $E(T | \theta) = \mu_T(\theta)$. This statistic, however, may not provide an unbiased estimate of θ itself, as when $\mu_T(\theta) \neq \theta$. Then

$$\mu_T(\theta) = \int \cdots \int T(y_1, \dots, y_N) L(y_1, \dots, y_N; \theta) dy_1 \cdots dy_N, \quad (\text{A.104})$$

where both T and $L(\theta)$ are functions of the observations. Differentiating with respect to θ and substituting (A.90) yields

$$\frac{d\mu_T(\theta)}{d\theta} = \mu'_T(\theta) = \int \cdots \int T \frac{d\ell(\theta)}{d\theta} L(\theta) dy_1 \cdots dy_N = E [TU(\theta)]. \quad (\text{A.105})$$

Since $E[U(\theta)] = 0$, then

$$\text{Cov}[TU(\theta)] = E \{ [T - \mu_T(\theta)] U(\theta) \} = E [TU(\theta)] = \mu'_T(\theta). \quad (\text{A.106})$$

We now apply the *Cauchy-Schwartz inequality*,

$$[E(AB)]^2 \leq E(A^2)E(B^2), \quad (\text{A.107})$$

so that

$$\begin{aligned} [\mu'_T(\theta)]^2 &= (E \{ [T - \mu_T(\theta)] U(\theta) \})^2 \\ &\leq E \left([T - \mu_T(\theta)]^2 \right) E \left([U(\theta)]^2 \right) = V(T)I(\theta). \end{aligned} \quad (\text{A.108})$$

Therefore,

$$V(T) \geq \frac{[\mu'_T(\theta)]^2}{I(\theta)} = \frac{[dE(T)/d\theta]^2}{E \left([d\ell(\theta)/d\theta]^2 \right)}. \quad (\text{A.109})$$

If $\mu'_T(\theta) = 1$, such as when T is unbiased for θ , then the variance of T is bounded by

$$V(T) \geq I(\theta)^{-1}. \quad (\text{A.110})$$

Similar developments apply to the multiparameter case where \mathbf{T} , $\mu'_T(\theta)$, and θ are each a p -vector. In this case, a lower bound can be determined for each element of the covariance matrix for \mathbf{T} (cf. Rao, 1973, p. 327). For the case where \mathbf{T} is unbiased for θ , the lower bound is provided by the inverse of the information as in (A.110).

From these results we can define an *efficient estimator* \mathbf{T} of the parameter θ as a *minimum variance unbiased estimator* (MVUE) such that $E(\mathbf{T}) = \theta$ and $V(\mathbf{T}) = I(\theta)^{-1}$.

A.6.5 Asymptotic Distribution of the Efficient Score and the MLE

To derive the asymptotic distribution of the maximum likelihood estimate, we first obtain that of the efficient score on which the estimates are based. To simplify matters, we consider only the single-parameter case where θ is a scalar.

Since the $\{y_i\}$ are *i.i.d.*, then likewise the scores $\{U_i(\theta)\} = \{d \log f(y_i; \theta) / d\theta\}$ are *i.i.d.* with expectation zero from (A.85). Thus,

$$V[U(\theta)] = \sum_{i=1}^N V[U_i(\theta)] = \sum_{i=1}^N E[U_i(\theta)^2] = I(\theta) = \sum_{i=1}^N E[-U'_i(\theta)] \quad (\text{A.111})$$

from the information equality. Further, since the total score $U(\theta)$ is the sum of *i.i.d.* random variables, and thus can be characterized as a sequence of partial sums as $n \rightarrow \infty$, then from the central limit theorem (A.14) it follows that

$$\frac{U(\theta)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, E[-U'_i(\theta)]), \quad (\text{A.112})$$

and asymptotically for large N

$$U(\theta) \xrightarrow{d} \mathcal{N}[0, I(\theta)], \quad (\text{A.113})$$

where $I(\theta) = NE[-U'_i(\theta)]$ for any random observation (the i th).

Now a Taylor's expansion of $U(\hat{\theta})$ about the value $U(\theta)$ yields asymptotically

$$U(\hat{\theta}) \cong U(\theta) + (\hat{\theta} - \theta)U'(\theta). \quad (\text{A.114})$$

By definition $U(\hat{\theta}) = 0$ and asymptotically

$$\sqrt{n}(\hat{\theta} - \theta) \cong -\frac{\sqrt{n}U(\theta)}{U'(\theta)} = \frac{U(\theta)/\sqrt{n}}{-U'(\theta)/n}. \quad (\text{A.115})$$

Since $-U'(\theta) = -\sum_{i=1}^n U'_i(\theta)$, from the law of large numbers it follows that

$$\frac{-U'(\theta)}{n} \xrightarrow{p} E[-U'_i(\theta)]. \quad (\text{A.116})$$

Thus, the numerator in (A.115) converges in distribution to a normal as in (A.112) above, whereas the denominator converges in probability to a constant. Applying Slutsky's theorem (A.44) we then obtain

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} \mathcal{N} \left[0, \frac{E[-U'_i(\theta)]}{(E[-U'_i(\theta)])^2} \right] \\ &\xrightarrow{d} \mathcal{N} [0, (E[-U'_i(\theta)])^{-1}]. \end{aligned} \quad (\text{A.117})$$

Since $I(\theta) = NE[-U'_i(\theta)]$, it follows that the large sample distribution of the *MLE* asymptotically is

$$\hat{\theta} - \theta \xrightarrow{d} \mathcal{N} [0, I(\theta)^{-1}]. \quad (\text{A.118})$$

An equivalent result also applies to the general p -vector parameter case such that the *MLE* asymptotically is distributed as

$$\hat{\theta} - \theta \xrightarrow{d} \mathcal{N} [\mathbf{0}, \mathbf{I}(\theta)^{-1}]. \quad (\text{A.119})$$

The large sample variance of the j th element of the parameter vector estimate, $V(\hat{\theta}_j)$, is obtained as the j th diagonal element of the inverse information matrix,

$$V(\hat{\theta}_j) = \Sigma_{\hat{\theta}j} = [\mathbf{I}(\theta)^{-1}]_{jj} = \mathbf{I}(\theta)^{jj}. \quad (\text{A.120})$$

A.6.6 Consistency and Asymptotic Efficiency of the MLE

From (A.118), the asymptotic variance of the *MLE* is

$$\lim_{n \rightarrow \infty} V(\hat{\theta}) = \lim_{n \rightarrow \infty} \frac{1}{nE[-U'_i(\theta)]} = \lim_{n \rightarrow \infty} I(\theta)^{-1} = 0, \quad (\text{A.121})$$

so that the distribution of the *MLE* as $n \rightarrow \infty$ is

$$(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 0), \quad (\text{A.122})$$

which is a degenerate normal distribution with variance zero. Thus, $\hat{\theta}$ converges in distribution to the constant θ , which, in turn, implies that $\hat{\theta} \xrightarrow{p} \theta$ and that the *MLE* is a consistent estimator for θ . A similar result applies in the multiparameter case. The *MLE*, however, is not unbiased with finite samples. In fact, the bias can be substantial when a set of parameters are to be estimated simultaneously; see Cox and Hinkley (1974).

Since the asymptotic variance of the *MLE* is $V(\hat{\theta}) = \mathbf{I}(\theta)^{-1}$, which is the Cramér-Rao lower bound for the asymptotic variance of a consistent estimator, then this also establishes that the *MLE* is asymptotically a minimum variance estimator, or is fully efficient.

From these results, an *asymptotically efficient estimator* \mathbf{T} of the parameter θ is one that is consistent, $\mathbf{T} \xrightarrow{p} \theta$, and for which the asymptotic variance equals $\mathbf{I}(\theta)^{-1}$.

A.6.7 Estimated Information

The expressions for the expected and observed information assume that the true value of θ is known. In practice, these quantities are estimated based on the maximum likelihood estimate $\hat{\theta}$ of θ . This leads to expressions for the estimated expected information and the estimated observed information.

The estimated Hessian is obtained from (A.97) evaluated at the value of the *MLE* is

$$\mathbf{H}(\hat{\theta}) = \mathbf{U}'(\hat{\theta}) = \mathbf{U}'(\theta)_{|\theta=\hat{\theta}}, \quad (\text{A.123})$$

from which the estimated observed information is obtained as $i(\hat{\theta}) = -\mathbf{H}(\hat{\theta})$. When the elements of the expected information exist in closed form in either (A.95), (A.96), (A.99), or (A.103), the estimated expected information, denoted as $\mathbf{I}(\hat{\theta})$, may be obtained by evaluating the resulting expressions at the values of the estimates; that is, $\mathbf{I}(\hat{\theta}) = \mathbf{I}(\theta)_{|\theta=\hat{\theta}}$.

Also, since the *MLE* $\hat{\theta}$ is consistent for θ , it follows from Slutsky's theorem that $i(\hat{\theta}) \xrightarrow{p} i(\theta)$ and that $\mathbf{I}(\hat{\theta}) \xrightarrow{p} \mathbf{I}(\theta)$ so that the estimated observed and expected information are also consistent estimates. This provides the basis for large sample confidence interval estimates of the parameters and tests of significance.

A.6.8 Invariance Under Transformations

Finally, an important property of the *MLE* is that it is invariant under one-to-one transformations of θ such as $g(\theta) = \log(\theta)$ for θ nonnegative, or $g(\theta) = \sqrt{\theta}$ for θ nonnegative, or $g(\theta) = e^\theta$ for $\theta \in \mathcal{R}$. In such cases, if $\hat{\theta}$ is the *MLE* of θ , then $g(\hat{\theta})$ is the *MLE* of $g(\theta)$. Note that this does not apply to functions such as θ^2 , which are not one-to-one.

Example A.9 Poisson-Distributed Counts

Consider the estimation and testing of the parameter of a Poisson distribution from a sample of N counts y_i , $i = 1, \dots, N$. Under the assumed model that the N counts

are independently and identically distributed as Poisson with rate parameter θ and probability distribution

$$f(y; \theta) = \frac{e^{-\theta} \theta^y}{y!} \quad y \geq 0, E(y) = \theta. \quad (\text{A.124})$$

Then the likelihood of the sample of N observations is

$$L(y_1, \dots, y_N; \theta) = \prod_{i=1}^N \frac{e^{-\theta} \theta^{y_i}}{y_i!} \quad (\text{A.125})$$

with

$$\begin{aligned} \ell(\theta) &= \log L(\theta) = \sum_{i=1}^N [-\theta + y_i \log(\theta) - \log(y_i!)] \\ &= -N\theta + \log(\theta) \sum_i y_i - \sum_i \log(y_i!). \end{aligned} \quad (\text{A.126})$$

Therefore,

$$U(\theta) = \frac{d\ell(\theta)}{d\theta} = \frac{\sum_i y_i}{\theta} - N, \quad (\text{A.127})$$

which, when set to zero, yields the *MLE*

$$\hat{\theta} = \frac{\sum_i y_i}{N}. \quad (\text{A.128})$$

The observed information function is

$$i(\theta) = -H(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2} = \frac{\sum_i y_i}{\theta^2}. \quad (\text{A.129})$$

Since $E[\sum_i y_i] = N\theta$, then the expected information function is

$$I(\theta) = E[-H(\theta)] = E\left[\frac{\sum_i y_i}{\theta^2}\right] = \frac{N}{\theta}. \quad (\text{A.130})$$

Therefore, the large sample variance of the estimate is

$$V(\hat{\theta}) = I(\theta)^{-1} = \theta/N \quad (\text{A.131})$$

and asymptotically

$$\hat{\theta} \xrightarrow{d} \mathcal{N}[\theta, \theta/N]. \quad (\text{A.132})$$

To construct confidence limits one can use either the estimated observed or the estimated expected information, which in this case are the same

$$i(\hat{\theta}) = \frac{N\hat{\theta}}{\hat{\theta}^2} = \frac{N}{\hat{\theta}} = I(\hat{\theta}) = \frac{N}{\hat{\theta}}, \quad (\text{A.133})$$

so that the estimated large sample variance is

$$\hat{V}(\hat{\theta}) = \frac{\hat{\theta}}{N}. \quad (\text{A.134})$$

Example A.10 Hospital Mortality

Consider the following hypothetical data from a sample of 10 hospitals. For each hospital, the following are the numbers of deaths among the first 1000 patient admissions during the past year: 8, 3, 10, 15, 4, 11, 9, 17, 6, 8. For this sample of 10 hospitals, $\sum_i y_i = 91$ and $\hat{\theta} = \hat{\mu} = 9.1$ per 1000 admissions. The estimated information is $I(\hat{\theta}) = 10/9.1 = 1.09890$, and the estimated standard error of the estimate is $\sqrt{9.1/10} = 0.95394$.

A.6.9 Independent But Not Identically Distributed Observations

Similar developments apply to the estimates of the parameters in a likelihood based on a sample of N observations that are statistically independent but are not identically distributed. In particular, suppose that the form of the distribution is the same for all observations, but the moments of the distribution of y_i are a function of covariates \mathbf{x}_i through a linear function of a parameter vector $\boldsymbol{\theta}$. Then the conditional distribution of $Y|\mathbf{x} \sim f(y; \mathbf{x}'\boldsymbol{\theta})$, where f is of the same form or family for all observations. Then the likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N f(y_i; \mathbf{x}'_i \boldsymbol{\theta}). \quad (\text{A.135})$$

Although the $\{y_i\}$ are no longer identically distributed, all of the above results still apply to the maximum likelihood estimates of the parameter vector $\boldsymbol{\theta}$ and the associated score statistics. The demonstration of these properties, however, is more tedious than that presented for *i.i.d.* observations. These properties are illustrated by the following example.

Example A.11 Homoscedastic Normal Errors Regression

Again, consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ of Section A.5.1. There, the ordinary least squares estimates of the parameters were obtained based only on the specification of the first and second moments of the distribution of the errors. Now consider the model where it is also assumed that the errors are independently and identically normally distributed with mean zero and constant variance, $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2)$. This is called the homoscedastic *i.i.d.* normal errors assumption. Therefore, conditioning on the covariate vector \mathbf{x}_i , then

$$Y|\mathbf{x}_i \sim N(\mathbf{x}'_i \boldsymbol{\theta}, \sigma_\varepsilon^2). \quad (\text{A.136})$$

Thus, conditionally, the $Y|\mathbf{x}_i$ are independently but not identically distributed.

The likelihood of a sample of N observations, each with response y_i and a covariate vector \mathbf{x}_i , is

$$L(y_1, \dots, y_N; \boldsymbol{\theta}, \sigma_\varepsilon^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp \left\{ -\frac{1}{2} \left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\theta}}{\sigma_\varepsilon} \right)^2 \right\}. \quad (\text{A.137})$$

Thus, the log likelihood, up to a constant term that does not depend on θ , is

$$\ell(\theta) = \sum_i \frac{-(y_i - \mathbf{x}'_i \theta)^2}{2\sigma_\epsilon^2} = \sum_i \frac{-(y_i^2 - 2y_i \mathbf{x}'_i \theta + (\mathbf{x}'_i \theta)^2)}{2\sigma_\epsilon^2}. \quad (\text{A.138})$$

Adopting matrix notation as in Section A.5.1, then

$$\ell(\theta) = \frac{-\mathbf{Y}'\mathbf{Y} + 2\theta'\mathbf{X}'\mathbf{Y} - \theta'\mathbf{X}'\mathbf{X}\theta}{2\sigma_\epsilon^2}. \quad (\text{A.139})$$

The total score vector then is

$$\mathbf{U}(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \frac{2\mathbf{X}'\mathbf{Y} - 2(\mathbf{X}'\mathbf{X})\theta}{2\sigma_\epsilon^2}, \quad (\text{A.140})$$

which yields the following estimating equation for θ when set to zero

$$(\mathbf{X}'\mathbf{X})\theta = \mathbf{X}'\mathbf{Y}. \quad (\text{A.141})$$

Therefore, the *MLE* of the coefficient vector θ equals the least squares estimate presented in (A.60). Since we condition on the covariate vectors $\{\mathbf{x}_i\}$, then the information function is

$$\mathbf{I}(\theta) = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right] = -E\left[\frac{-\mathbf{X}'\mathbf{X}}{\sigma_\epsilon^2}\right] = \frac{\mathbf{X}'\mathbf{X}}{\sigma_\epsilon^2} \quad (\text{A.142})$$

and the large sample variance of the estimates is

$$V(\hat{\theta}) = \mathbf{I}(\theta)^{-1} = (\mathbf{X}'\mathbf{X})^{-1} \sigma_\epsilon^2, \quad (\text{A.143})$$

which equals the variance of the least squares estimates presented in (A.63). Finally, the vector of the *MLEs* of the coefficients are asymptotically normally distributed as

$$\hat{\theta} \stackrel{d}{\approx} \mathcal{N}\left[\theta, (\mathbf{X}'\mathbf{X})^{-1} \sigma_\epsilon^2\right]. \quad (\text{A.144})$$

A.7 TESTS OF SIGNIFICANCE

The above developments lead to three different approaches to conducting a large sample test for the values of the parameter vector of the assumed model, or for elements of the parameter vector: the Wald test, the likelihood ratio test, and the efficient score test. These tests are described in terms of a p -vector of parameters θ .

A.7.1 Wald Tests

Of the various types of tests, the Wald test requires the least computational effort, and thus is the most widely used. It can be applied to any vector of statistics that are

asymptotically distributed as multivariate normal with a consistent estimate of the covariance matrix of the parameter estimates, $\widehat{\Sigma}_{\widehat{\theta}}$, such as for parameter estimates provided by least squares or maximum likelihood estimation in a given model. For ordinary least squares estimates, $\widehat{\Sigma}_{\widehat{\theta}}$ is provided by (A.65) using a consistent estimate of the error variance, and for weighted least squares estimates by (A.72), using a consistent estimate of the error covariance structure. For maximum likelihood estimates, the large sample covariance matrix can be estimated consistently as $\widehat{\Sigma}_{\widehat{\theta}} = \mathbf{I}(\widehat{\theta})^{-1} \xrightarrow{P} \Sigma_{\widehat{\theta}}$. The following assumes that the distribution is not degenerate or that the covariance matrix is of full rank p .

A.7.1.1 Elementwise Tests Consider that we wish to test $H_{0j}: \theta_j = \theta_0$ for the j th element of the vector θ versus the alternative hypothesis $H_{1j}: \theta_j \neq \theta_0$ based on the sample estimate $\widehat{\theta}_j$ and estimated variance $\widehat{V}(\widehat{\theta}_j) = \widehat{\Sigma}_{\widehat{\theta}_{jj}}$. From the important work of Wald (1943) and others, a large sample test of H_{0j} is provided by the statistic

$$X_{W_j}^2 = \frac{(\widehat{\theta}_j - \theta_0)^2}{\widehat{V}(\widehat{\theta}_j)}, \quad (\text{A.145})$$

that is asymptotically distributed as chi-square on 1 df . In the case of maximum likelihood estimation, $\widehat{V}(\widehat{\theta}_j) = [\mathbf{I}(\widehat{\theta})^{-1}]_{jj} = \mathbf{I}(\widehat{\theta})^{jj}$.

A.7.1.2 Composite Test Now assume that we wish to test $H_0: \theta = \theta_0$ for the complete p -vector versus the alternative $H_1: \theta \neq \theta_0$. The Wald large sample test is a T^2 -like test statistic:

$$X_W^2 = (\widehat{\theta} - \theta_0)' \widehat{\Sigma}_{\widehat{\theta}}^{-1} (\widehat{\theta} - \theta_0), \quad (\text{A.146})$$

that is asymptotically distributed as chi-square on p df .

A.7.1.3 Test of a Linear Hypothesis Sometimes we wish to test a hypothesis about the value of a linear combination or contrast of the elements of θ . In this case we wish to test a null hypothesis of the form

$$H_{0C}: \mathbf{C}'\theta = \mathbf{K}, \quad (\text{A.147})$$

where \mathbf{C}' is an $s \times p$ matrix of rank s ($\leq p$) and \mathbf{K} is the $s \times 1$ solution vector. The i th row of \mathbf{C}' is a linear combination of the elements of θ that yields the solution K_i specified by the i th element of \mathbf{K} . The Wald test of the hypothesis H_{0C} is then provided by the large sample T^2 -like test statistic

$$X_W^2 = (\mathbf{C}'\widehat{\theta} - \mathbf{K})' [\mathbf{C}'\widehat{\Sigma}_{\widehat{\theta}}\mathbf{C}]^{-1} (\mathbf{C}'\widehat{\theta} - \mathbf{K}), \quad (\text{A.148})$$

that is asymptotically distributed as chi-square on s df .

A common special case is where we wish to test a hypothesis regarding the values of a subset of the parameter vector θ . In the latter case, let the parameter vector be

partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 // \boldsymbol{\theta}_2)$, the two subvectors consisting of r and s elements, respectively, $r + s = p$. Then we wish to test $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{0(r)}$ irrespective of the values of $\boldsymbol{\theta}_2$, where $\boldsymbol{\theta}_{0(r)}$ is the r -element subvector of elements $\boldsymbol{\theta}_0$ specified under the null hypothesis. This hypothesis can be expressed as a simple linear contrast on the elements of $\boldsymbol{\theta}$ of the form H_{0C} : $\mathbf{C}'\boldsymbol{\theta} = \boldsymbol{\theta}_{0(r)}$, where \mathbf{C}' is a $r \times p$ matrix with elements

$$\mathbf{C}' = [\mathbf{I}_r \parallel \mathbf{0}_{r \times s}]_{r \times p}, \quad (\text{A.149})$$

meaning a $r \times r$ identity matrix augmented by a $r \times s$ matrix of zeros. The Wald test of the hypothesis H_{0C} is then provided by

$$X_W^2 = (\mathbf{C}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{0(r)})' [\mathbf{C}'\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} \mathbf{C}]^{-1} (\mathbf{C}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{0(r)}), \quad (\text{A.150})$$

that is asymptotically distributed as chi-square on r df.

The most common application is the test of significance of the set of regression coefficients in a regression model, not including the intercept. In a model with p -covariates, the parameter vector $\boldsymbol{\theta} = (\alpha \ \beta_1 \ \cdots \ \beta_p)^T$ and we wish to test the significance of the model for which the null hypothesis is $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$. The contrast matrix of this *model chi-square test* is $\mathbf{C}' = [\mathbf{0}_{p \times 1} \parallel \mathbf{I}_p]$.

In most cases, a Wald test uses the variance of the estimates $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ evaluated under the alternative hypothesis, not under the null hypothesis of interest. Thus, these tests will not be as efficient as would comparable tests for which the covariance matrix of the estimates is evaluated under the null hypothesis. However, all such tests are asymptotically equivalent because the covariance matrix estimated without restriction under the alternative hypothesis is still consistent for the true covariance matrix when the null hypothesis is true. In some cases, however, it is possible to compute a Wald test using the variance estimated under the null hypothesis, such as $\hat{\Sigma}_{\boldsymbol{\theta}_0} = \mathbf{I}(\boldsymbol{\theta}_0)$, in which case the test statistic comparable to (A.146) is

$$X_{W_0}^2 = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \hat{\Sigma}_{\boldsymbol{\theta}_0}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \quad (\text{A.151})$$

Likewise the contrast test in (A.148) would employ $\Sigma_{\boldsymbol{\theta}_0}$ and the test of an individual parameter in (A.145) would employ $V_0(\hat{\theta}_j) = \hat{\Sigma}_{\hat{\boldsymbol{\theta}}_{0,jj}}$. This approach is preferable because the size of the test with large samples will more closely approximate the desired Type I error probability level.

Finally, in the case that the underlying distribution is a degenerate multivariate normal of rank $r < p$, such as for a multinomial distribution, a composite test of the parameter vector $\boldsymbol{\theta}$ can be constructed using an appropriate subvector of $\boldsymbol{\theta}$. This in turn can be expressed as a contrast test (A.148), where $\mathbf{C}'\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} \mathbf{C}$ is nonsingular of rank r .

A.7.2 Likelihood Ratio Tests

A.7.2.1 Composite Test Another type of test is the likelihood ratio test, which is the uniformly most powerful test of $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1: \boldsymbol{\theta} = \boldsymbol{\theta}_1$ when both

the null and alternative hypothesis values $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ are completely specified. For a test against the omnibus alternative hypothesis $H_1: \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, it is necessary that the value of the parameter under the alternative be estimated from the data. When $\boldsymbol{\theta}$ is a p -vector, the null likelihood is

$$L(\boldsymbol{\theta}_0) = \prod_{i=1}^N f(y_i; \boldsymbol{\theta})_{|\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \quad (\text{A.152})$$

Under the alternative hypothesis, the likelihood function is estimated using the vector of *MLEs* $\widehat{\boldsymbol{\theta}}$,

$$L(\widehat{\boldsymbol{\theta}}) = \prod_{i=1}^N f(y_i; \boldsymbol{\theta})_{|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \quad (\text{A.153})$$

Then the likelihood ratio test is

$$X_{L(p)}^2 = -2 \log \left[\frac{L(\boldsymbol{\theta}_0)}{L(\widehat{\boldsymbol{\theta}})} \right] = 2 \log L(\widehat{\boldsymbol{\theta}}) - 2 \log L(\boldsymbol{\theta}_0), \quad (\text{A.154})$$

which is asymptotically distributed as χ_p^2 on p *df* under the null hypothesis.

In general, the quantity $-2 \log L(\boldsymbol{\theta})$ is analogous to the *SSE* in a simple linear regression model, so that $X_{L(p)}^2$ is analogous to the reduction in *SSE* associated with the addition of the parameter vector $\boldsymbol{\theta}$ to the model.

A.7.2.2 Test of a Subhypothesis The most common application of a likelihood ratio test is to test nested subhypotheses that include the model test in a regression model and tests of individual elements. As for the Wald test, assume that the p -vector $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 // \boldsymbol{\theta}_2)$ of r and s elements, respectively, where we wish to test $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{0(r)}$ versus the alternative hypothesis $H_1: \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{0(r)}$. This requires that we compare the likelihoods from two fitted models, that using the complete p -vector $\boldsymbol{\theta}$ versus that using only the complement of the subvector to be tested, in this case the s -vector $\boldsymbol{\theta}_2$. The likelihood ratio test then is

$$X_{L(r)}^2 = -2 \log \left[\frac{L(\widehat{\boldsymbol{\theta}}_2)}{L(\widehat{\boldsymbol{\theta}})} \right] = 2 \log L(\widehat{\boldsymbol{\theta}}) - 2 \log L(\widehat{\boldsymbol{\theta}}_2), \quad (\text{A.155})$$

that is asymptotically distributed as χ_r^2 on r *df* under the null subhypothesis. This can also be viewed as the difference between two independent chi-square statistics relative to the null likelihood $L(\boldsymbol{\theta}_0)$, such as

$$X_{L(r)}^2 = -2 \log \left[\frac{L(\boldsymbol{\theta}_0)}{L(\widehat{\boldsymbol{\theta}})} \right] - \left(-2 \log \left[\frac{L(\boldsymbol{\theta}_0)}{L(\widehat{\boldsymbol{\theta}}_2)} \right] \right) = X_{L(p)}^2 - X_{L(s)}^2 \quad (\text{A.156})$$

with degrees of freedom equal to $r = p - s$.

To test $H_{0j}: \theta_j = \theta_0$ versus $H_{1j}: \theta_j \neq \theta_0$ for the j th element of $\boldsymbol{\theta}$ requires that one evaluate the difference between the log likelihoods for the complete p -vector $\boldsymbol{\theta}$ and for the subset with the j th element excluded. Thus, the computation of the likelihood ratio tests for the elements of a model can be more tedious than the Wald tests. However, the likelihood ratio tests, in general, are preferred because they have greater efficiency or power.

A.7.3 Efficient Scores Test

Rao (1963), among others, proposed that the efficient score vector be used as the basis for a statistical test for the assumed parameters. From (A.86), under a hypothesis regarding the true values of the parameter vector θ such as $H_0: \theta = \theta_0$, it follows that $E[U(\theta_0)] = 0$. If the data agree with the tested hypothesis, then the score statistic evaluated at θ_0 should be close to zero. If the data do not agree with H_0 , then we expect $U(\theta_0)$ to differ from zero.

A.7.3.1 Composite Test To test a composite hypothesis $H_0: \theta = \theta_0$ regarding the elements of the p -vector θ , from (A.113) under H_0 asymptotically

$$U(\theta_0) \stackrel{d}{\approx} N[0, I(\theta_0)]. \quad (\text{A.157})$$

Thus, a large sample test of H_0 versus the alternative $H_1: \theta \neq \theta_0$ is provided by

$$X_S^2 = U(\theta_0)' I(\theta_0)^{-1} U(\theta_0), \quad (\text{A.158})$$

which is asymptotically distributed as χ^2 on p df. Note that in order to conduct a score test regarding the complete parameter vector θ does not require that the *MLE* $\hat{\theta}$ or the estimated information $I(\hat{\theta})$ be computed because the score equation and the expected information are evaluated under the null hypothesis parameter values.

A.7.3.2 Test of a Subhypothesis: $C(\alpha)$ Tests Score tests may also be constructed for subhypotheses regarding elements of the vector θ . Such tests were originally described by Neyman (1959), who referred to such tests as $C(\alpha)$ tests, α designating the nuisance parameters. Most of the score tests of subhypotheses considered herein involve a test for the value of one of two parameters. Thus, we first consider the case of a two-parameter vector $\theta = (\alpha \beta)^T$, where we wish to test $H_{\beta_0}: \beta = \beta_0$. This H_{β_0} implies the joint null hypothesis $H_{\theta_0}: \theta = \theta_0 = (\alpha_0 \beta_0)^T$, where the value of α is unrestricted. Under this hypothesis the bivariate score vector is

$$U(\theta_0) = \left[U(\theta)_\alpha \ U(\theta)_\beta \right]_{\beta=\beta_0}^T. \quad (\text{A.159})$$

However, because the hypothesis to be tested makes no restrictions on the value of the nuisance parameter α , it is necessary to estimate α under the restriction that $\beta = \beta_0$. The *MLE* of α , designated as $\hat{\alpha}_0$, is obtained as the solution to the estimating equation $U(\theta)_\alpha = 0$ evaluated under the null hypothesis $H_{\beta_0}: \beta = \beta_0$. Thus, the estimated parameter vector under H_{β_0} is $\hat{\theta}_0 = (\hat{\alpha}_0 \beta_0)^T$.

The resulting score vector can be expressed as

$$U(\hat{\theta}_0) = \left[U(\theta)_\alpha \ U(\theta)_\beta \right]_{\hat{\alpha}_0, \beta_0}^T. \quad (\text{A.160})$$

By definition, the first element of the score vector is

$$U(\hat{\theta}_0)_\alpha = [U(\theta)_\alpha]_{\hat{\alpha}_0, \beta_0} = 0, \quad (\text{A.161})$$

since $\hat{\alpha}_0$ is the value that satisfies this equality. However, the second element,

$$U(\hat{\theta}_0)_\beta = [U(\theta)_\beta]_{|\hat{\alpha}_0, \beta_0}, \quad (\text{A.162})$$

may not equal 0 and, in fact, will only equal 0 when the solution to the score equation for β is $\hat{\beta} = \beta_0$. Note that we must actually solve for the nuisance parameter α under the restriction that $\beta = \beta_0$, in order to evaluate the score statistic for β under the null hypothesis.

Therefore, the bivariate score vector is

$$U(\hat{\theta}_0) = \begin{bmatrix} 0 & U(\hat{\theta}_0)_\beta \end{bmatrix}^T, \quad (\text{A.163})$$

which is a random variable augmented by a constant 0. The corresponding estimated information function is

$$I(\hat{\theta}_0) = I(\theta)|_{\theta=\hat{\theta}_0}, \quad (\text{A.164})$$

meaning that the elements are evaluated at the values $\hat{\alpha}_0$ and β_0 . Then the score test is the quadratic form

$$\begin{aligned} X_{S(1)}^2 &= U(\hat{\theta}_0)' I(\hat{\theta}_0)^{-1} U(\hat{\theta}_0) \\ &= \begin{bmatrix} 0 & U(\hat{\theta}_0)_\beta \end{bmatrix} \begin{bmatrix} I(\hat{\theta}_0)^\alpha & I(\hat{\theta}_0)^{\alpha\beta} \\ I(\hat{\theta}_0)^{\beta\alpha} & I(\hat{\theta}_0)^\beta \end{bmatrix} \begin{bmatrix} 0 \\ U(\hat{\theta}_0)_\beta \end{bmatrix} \\ &= U(\hat{\theta}_0)_\beta' I(\hat{\theta}_0)^\beta U(\hat{\theta}_0)_\beta, \end{aligned} \quad (\text{A.165})$$

which is asymptotically distributed as χ_1^2 . A Wald test for such hypotheses is easy to compute, but score tests have the advantage that the variance of the test is evaluated under the null hypothesis for the parameters of interest.

As a special case, this approach also includes a test of an individual element of the parameter vector, such as a test of $H_0: \theta_j = \theta_0$ versus $H_1: \theta_j \neq \theta_0$ for the j th element of θ . This score test is a bit more tedious than a Wald test because the terms that are not restricted by the hypothesis (the $\hat{\alpha}$) must be estimated and included in the evaluation of the score statistics for the parameters that are restricted, and included in the computation of the estimated information function and its inverse. If we wish to conduct score tests for multiple elements of the parameter vector separately, then a separate model must be fit under each hypothesis.

For the two-parameter example, to also test the $H_{\alpha_0}: \alpha = \alpha_0$ with no restrictions on β requires that we refit the model to obtain estimates of $\hat{\theta}_0 = (\alpha_0 \ \hat{\beta}_0)^T$ and to then compute the score test as $X_S^2 = U(\hat{\theta}_0)'_\alpha I(\hat{\theta}_0)^\alpha U(\hat{\theta}_0)_\alpha$.

In the more general case, the p -vector θ may be partitioned as $\theta = (\theta_1 // \theta_2)$ of r and s elements, respectively, as for the Wald test, and we wish to test $H_0: \theta_1 =$

$\theta_{0(r)}$. Then for $\theta_0 = (\theta_{0(r)} // \theta_2)$ the MLE under the tested hypothesis is $\hat{\theta}_0 = (\theta_{0(r)} // \hat{\theta}_2)$ with corresponding score vector $\mathbf{U}(\hat{\theta}_0) = [\mathbf{U}(\hat{\theta}_0)_{\theta_1} \ \mathbf{U}(\hat{\theta}_0)_{\theta_2}]^T$. The score test is

$$\begin{aligned} X_{S(r)}^2 &= \mathbf{U}(\hat{\theta}_0)' \mathbf{I}(\hat{\theta}_0)^{-1} \mathbf{U}(\hat{\theta}_0) \\ &= [\mathbf{U}(\hat{\theta}_0)_{\theta_1} \ \mathbf{U}(\hat{\theta}_0)_{\theta_2}] \mathbf{I}(\hat{\theta}_0)^{-1} [\mathbf{U}(\hat{\theta}_0)_{\theta_1} \ \mathbf{U}(\hat{\theta}_0)_{\theta_2}]^T. \end{aligned} \quad (\text{A.166})$$

However, by definition, $\mathbf{U}(\hat{\theta}_0)_{\theta_2} = [\mathbf{U}(\theta)_{\theta_2}]_{|\hat{\theta}_2, \theta_{0(r)}} = \mathbf{0}$, so that

$$X_{S(r)}^2 = [\mathbf{U}(\hat{\theta}_0)_{\theta_1}]^T \mathbf{I}(\hat{\theta}_0)^{\theta_1} \mathbf{U}(\hat{\theta}_0)_{\theta_1}. \quad (\text{A.167})$$

In this expression, $\mathbf{I}(\hat{\theta}_0)^{\theta_1}$ is the upper left $r \times r$ submatrix of $\mathbf{I}(\hat{\theta}_0)^{-1}$ that can be obtained from the expression for the inverse of a patterned matrix in (A.3).

A.7.3.3 Relative Efficiency Versus the Likelihood Ratio Test Score tests are called *efficient score tests* because they can be shown to be asymptotically fully efficient with power approaching that of the UMP likelihood ratio test. For illustration, consider a test of $H_0: \theta = \theta_0$, where θ_0 is a scalar parameter, versus a local alternative hypothesis $H_1: \theta_n = \theta_0 + \delta/\sqrt{n}$ with some fixed quantity δ such that $\theta_n \rightarrow \theta_0$ as $n \rightarrow \infty$. Then the likelihood ratio statistic is

$$X_L^2 = -2 \log \left[\frac{L(\theta_0)}{L(\theta_0 + \delta/\sqrt{n})} \right] = 2 \log L \left(\theta_0 + \frac{\delta}{\sqrt{n}} \right) - 2 \log L(\theta_0), \quad (\text{A.168})$$

which is asymptotically distributed as χ_1^2 under H_0 . Now consider a Taylor's expansion of $\log L(\theta_0 + \delta/\sqrt{n})$ about the value θ_0 . Then

$$2 \log L \left(\theta_0 + \frac{\delta}{\sqrt{n}} \right) = 2 \log L(\theta_0) + \left(\frac{\delta}{\sqrt{n}} \right) \left[\frac{d[2 \log L(\theta_0)]}{d\theta_0} \right] + R_2, \quad (\text{A.169})$$

where R_2 is the remainder involving the term $(\delta/\sqrt{n})^2$ that vanishes in the limit. Therefore, asymptotically,

$$X_L^2 \cong (\delta/\sqrt{n}) 2U(\theta_0) \propto U(\theta_0) \quad (\text{A.170})$$

and the efficient score test based on the score function $U(\theta_0)$ is a locally optimum test statistic.

In general, therefore, a likelihood ratio test or score test is preferred to a Wald test unless the latter is computed using the variance estimated under the tested hypothesis as in (A.151). Asymptotically, however, it has also been shown that the Wald test is approximately equal to the likelihood ratio test, since under the null hypothesis, the variance estimated under no restrictions converges to the true null variance (cf. Cox and Hinkley, 1974).

Example A.12 Poisson Counts

For the above example of hospital mortality, suppose that we wish to test the hypothesis that the mortality rate in these ten hospitals was 12 deaths per 1000 admissions, or $H_0: \theta = \theta_0 = 12$ deaths/1000 admissions. The Wald test using the estimated information is $X_W^2 = (\hat{\theta} - \theta_0)^2 I(\hat{\theta}) = (9.1 - 12)^2 (1.0989) = 9.24$, with $p < 0.0024$.

To compute the likelihood ratio test of $H_0: \theta_0 = 12$ deaths/1000 admissions, the null log likelihood up to an additive constant is $\ell(\theta_0) = [-N\theta_0 + \log(\theta_0) \sum_i y_i] = ([\log(12)](91) - (10)(12)) = 106.127$. The MLE is $\hat{\theta} = 9.1$ and the corresponding log likelihood, up to a constant, is $\ell(\hat{\theta}) = [-N\hat{\theta} + \log(\hat{\theta}) \sum_i y_i] = ([\log(9.1)](91) - (10)(9.1)) = 109.953$. Thus, the likelihood ratio test is $X_L^2 = 2(109.953 - 106.127) = 7.653$, with $p < 0.0057$.

The score test is computed using the score $U(\theta_0) = \sum_i y_i / \theta_0 - N = (91/12) - 10 = 2.4167$ and the information function evaluated under the hypothesis $I(\theta_0) = N/\theta_0 = 10/12 = 0.8333$. Therefore, the score test is $X_S^2 = (2.4167)^2 / 0.8333 = 7.008$ with $p < 0.0081$.

For this example, the Wald test statistic is greater than both the likelihood ratio test statistic and the score test statistic. Asymptotically, all three tests are equivalent, however, with finite samples the three tests will differ, sometimes the Wald test being greater than the likelihood ratio test, sometimes less. However, since the Wald test employs the variance estimated under the alternative, the size or Type I error probability of the Wald test may be affected, and any apparent increase in power may be associated with an inflation in the test size. In general, therefore, either a likelihood ratio or score test is preferred to a Wald test that uses the estimated alternative variance.

However, the Wald test can also be computed using the variance evaluated under the null hypothesis that is obtained as the inverse of the information evaluated under the null hypothesis that forms the basis of the score test, rather than using the information evaluated under the alternative. In this example, the null-variance Wald test is $X_{W_0}^2 = (\hat{\theta} - \theta_0)^2 I(\theta_0) = (9.1 - 12)^2 (0.8333) = 7.008$, which, in this case, equals the score test. However, this will not be the case in general, especially in the multiparameter case.

Example A.13 Normal Errors Model Score Test

To illustrate the construction of a $C(\alpha)$ test, consider the test of the difference between two group means in the normal errors linear model of Example A.11. The covariate vector for the i th observation is $\mathbf{x}_i = (1 \ x_{i1})$, where $x_{i1} = 0$ if a member of the control group of size n_0 , and 1 if the treated group of size n_1 , with respective means μ_0 and μ_1 . Then the parameter vector $\boldsymbol{\theta} = (\alpha \ \beta)^T$, where $\alpha = \mu_0$ and $\beta = \mu_1 - \mu_0$. We thus wish to derive a score test of the hypothesis $H_0: \mu_1 = \mu_0$ or that $\beta = \beta_0 = 0$, for which $\boldsymbol{\theta}_0 = (\alpha \ \beta_0)^T$. Then from (A.140) it is readily shown that

$$\mathbf{U}(\boldsymbol{\theta}) = \begin{bmatrix} U(\boldsymbol{\theta})_\alpha \\ U(\boldsymbol{\theta})_\beta \end{bmatrix} = \frac{1}{\sigma_\varepsilon^2} \begin{bmatrix} \sum_i y_i - N\alpha - n_1\beta \\ \sum_i x_{i1}y_i - n_1\alpha - n_1\beta \end{bmatrix}, \quad (\text{A.171})$$

where $\sum_i x_i y_i$ is the sum of Y within group 1. Evaluating $U(\boldsymbol{\theta})_\alpha$ under H_0 yields

$$U(\boldsymbol{\theta}_0)_\alpha = \frac{\sum_i y_i - N\alpha}{\sigma_\varepsilon^2}, \quad (\text{A.172})$$

that when set equal to zero yields $\hat{\alpha}_0 = \sum_i y_i/N = \bar{y}_0$. Substituting into $U(\boldsymbol{\theta})_\beta$ evaluated under H_0 then yields

$$U(\boldsymbol{\theta}_0)_\beta = \frac{\sum_i x_i y_i - N\hat{\alpha}_0}{\sigma_\varepsilon^2} = \frac{n_0 n_1 (\bar{y}_1 - \bar{y}_0)}{N \sigma_\varepsilon^2}. \quad (\text{A.173})$$

Then, evaluating (A.142) yields

$$\mathbf{I}(\boldsymbol{\theta}_0) = \frac{1}{\sigma_\varepsilon^2} \begin{bmatrix} N & n_1 \\ n_1 & n_1 \end{bmatrix} \quad \text{and} \quad \mathbf{I}(\boldsymbol{\theta}_0)^{-1} = \frac{\sigma_\varepsilon^2}{N n_1 - n_1^2} \begin{bmatrix} n_1 & -n_1 \\ -n_1 & N \end{bmatrix}, \quad (\text{A.174})$$

where

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^\beta = \frac{N \sigma_\varepsilon^2}{N n_1 - n_1^2} = \frac{N \sigma_\varepsilon^2}{n_0 n_1}. \quad (\text{A.175})$$

From (A.167), the score test is provided by

$$X_{S(\beta)}^2 = \left[U(\hat{\boldsymbol{\theta}}_0)_\beta \right]^2 \mathbf{I}(\hat{\boldsymbol{\theta}}_0)^\beta = \frac{n_0 n_1 (\bar{y}_1 - \bar{y}_0)^2}{N \sigma_\varepsilon^2} \quad (\text{A.176})$$

that is distributed as χ_1^2 on 1 *df*. It is also the square of the large sample *Z*-test for the difference of two means with known or estimated error variance σ_ε^2 .

A.8 EXPLAINED VARIATION

One of the objectives of any model is to describe factors that account for variation among observations. It is also useful, therefore, to quantify the proportion of the total variation in the data that is explained by the model and its components. In ordinary multiple regression, as in Section A.5.1, this is expressed as $R^2 = SS(\text{model})/SS_y$, where $SS(\text{model})$ is the sum of squares for variation in Y explained by the model and SS_y is the total sum of squares of Y . The $SS(\text{model})$ is also obtained as $SS_y - SSE$, where SSE is the residual sum of squares of errors not explained by the model. Analogous measures can be derived for other models with different error structures other than the homoscedastic normal errors assumed in the multiple regression model. Korn and Simon (1991) present an excellent review of this area.

In general, let $\mathcal{L}[y, a(\mathbf{x})]$ represent the loss incurred by predicting y using a *prediction function* $a(\mathbf{x})$ that is a function of a covariate vector \mathbf{x} . Then the expected loss associated with $a(\mathbf{x})$ is

$$E(\mathcal{L}[y, a(\mathbf{x})]) = \int \mathcal{L}[y, a(\mathbf{x})] dF(\mathbf{x}, y), \quad (\text{A.177})$$

where $F(\mathbf{x}, y)$ is the joint cdf of \mathbf{X} and Y . Usually, we select $a(\mathbf{x}) = \tilde{y}(\mathbf{x})$, which is defined as the prediction function such that $E[\mathcal{L}]$ is minimized. Then the expected loss using the prediction function $\tilde{y}(\mathbf{x})$ is denoted as

$$D_{\mathcal{L}}(\mathbf{x}) = E(\mathcal{L}[y, \tilde{y}(\mathbf{x})]). \quad (\text{A.178})$$

This is then contrasted with the expected loss D_0 using an unconditional prediction function \tilde{y}_0 that does not depend on any covariates and which is constant for all observations, where

$$D_{\mathcal{L}(0)} = E(\mathcal{L}[y, \tilde{y}_0]). \quad (\text{A.179})$$

The resulting fraction of expected loss explained by the prediction function then is

$$\rho_{\mathcal{L}}^2 = \frac{D_{\mathcal{L}(0)} - D_{\mathcal{L}}(\mathbf{x})}{D_{\mathcal{L}(0)}}. \quad (\text{A.180})$$

A.8.1 Squared Error Loss

The most common loss function used to assess the adequacy of predictions, and to derive measures of explained variation, is squared error loss

$$\mathcal{L}(y, a(\mathbf{x})) = [y - a(\mathbf{x})]^2 = \varepsilon(\mathbf{x})^2, \quad (\text{A.181})$$

for any prediction function $a(\mathbf{x})$. The expected squared error loss then is

$$E(\mathcal{L}) = E[y - a(\mathbf{x})]^2 = \int [y - a(\mathbf{x})]^2 dF(\mathbf{x}, y). \quad (\text{A.182})$$

From ordinary least squares, $E(\mathcal{L})$ is minimized using the prediction function

$$\tilde{y}(\mathbf{x}) = E(Y|\mathbf{x}) = \mu_{y|\mathbf{x}} \quad (\text{A.183})$$

with expected squared error loss (ε^2),

$$D_{\varepsilon^2}(\mathbf{x}) = E[y - \mu_{y|\mathbf{x}}]^2 = E[V(Y|\mathbf{x})]. \quad (\text{A.184})$$

Unconditionally, or not using any covariate information, then

$$\tilde{y}_0 = E(Y) = \mu_y, \quad (\text{A.185})$$

and the expected loss is

$$D_{\varepsilon^2(0)} = E(y - \mu_y)^2 = V(Y) = \sigma_y^2. \quad (\text{A.186})$$

Thus, the fraction of squared error loss (ε^2) explained by the covariates \mathbf{x} in the model is

$$\rho_{\varepsilon^2}^2 = \frac{D_{\varepsilon^2(0)} - D_{\varepsilon^2}(\mathbf{x})}{D_{\varepsilon^2(0)}} = \frac{V(Y) - E[V(Y|\mathbf{x})]}{V(Y)}. \quad (\text{A.187})$$

Then, using (A.6), we can partition $V(Y) = D_{\varepsilon^2(0)}$ as

$$\begin{aligned} V(Y) &= E[V(Y|\mathbf{x})] + V[E(Y|\mathbf{x})] \\ &= E[D_{\varepsilon^2}(\mathbf{x})] + V[E(Y|\mathbf{x})], \end{aligned} \quad (\text{A.188})$$

so that

$$\rho_{\varepsilon^2}^2 = \frac{V[E(Y|\mathbf{x})]}{V(Y)} = \frac{E[E(Y|\mathbf{x})] - E(Y)}{E[Y - E(Y)]^2}. \quad (\text{A.189})$$

This can be estimated as

$$R_{\varepsilon^2}^2 = \frac{\widehat{V}[E(Y|\mathbf{x})]}{\widehat{\sigma}_y^2} = \frac{\widehat{D}_{\varepsilon^2(0)} - \widehat{D}_{\varepsilon^2}(\mathbf{x})}{\widehat{D}_{\varepsilon^2(0)}}, \quad (\text{A.190})$$

where

$$\widehat{V}[E(Y|\mathbf{x})] = \frac{1}{N} \sum_i \left[\widehat{E}(y_i|\mathbf{x}_i) - \widehat{E}(y_i|\mathbf{x}_i) \right]^2 = \frac{1}{N} \sum_i \left[\widehat{E}(y_i|\mathbf{x}_i) - \bar{y} \right]^2, \quad (\text{A.191})$$

\bar{y} being the sample mean of Y .

Example A.14 Multiple Linear Regression Model

In the ordinary multiple regression model we assume that $y = \mathbf{x}'\boldsymbol{\theta} + \varepsilon$, where $\mathbf{x}'\boldsymbol{\theta} = \alpha + x_1\beta_1 + \dots + x_p\beta_p$, and where the errors are independently and identically distributed with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma_\varepsilon^2$. Then $E(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\theta} = \tilde{y}(\mathbf{x})$ and the expected squared error loss is

$$D_{\varepsilon^2}(\mathbf{x}) = E[V(Y|\mathbf{x})] = \sigma_\varepsilon^2. \quad (\text{A.192})$$

Since we assume that $\boldsymbol{\theta}$ is known, the term $V[E(Y|\mathbf{x})]$ equals $V[\mathbf{x}'\boldsymbol{\theta}]$ with respect to the distribution of the covariate vector \mathbf{X} with covariance matrix $\Sigma_{\mathbf{x}}$. Thus,

$$V[E(Y|\mathbf{x})] = \boldsymbol{\theta}'\Sigma_{\mathbf{x}}\boldsymbol{\theta} \quad (\text{A.193})$$

and

$$\rho_{\varepsilon^2}^2 = \frac{\boldsymbol{\theta}'\Sigma_{\mathbf{x}}\boldsymbol{\theta}}{\boldsymbol{\theta}'\Sigma_{\mathbf{x}}\boldsymbol{\theta} + \sigma_\varepsilon^2}, \quad (\text{A.194})$$

(Helland, 1987). Note that the first element of \mathbf{x} is a constant, so that the first row and column of $\Sigma_{\mathbf{x}}$ are 0 and the intercept makes no contribution to this expression.

Now, given the least squares estimate of the parameter vector, $\widehat{\boldsymbol{\theta}}$, the numerator of (A.194) can be estimated as

$$\widehat{V}[E(Y|\mathbf{x})] = \frac{1}{N} \sum_i \left(\mathbf{x}'_i \widehat{\boldsymbol{\theta}} - \bar{\mathbf{x}}' \widehat{\boldsymbol{\theta}} \right)^2 = \frac{1}{N} \sum_i (\widehat{y}_i - \bar{y})^2, \quad (\text{A.195})$$

since $\widehat{y}_i = \mathbf{x}'_i \widehat{\boldsymbol{\theta}}$ and $\bar{y} = \bar{\mathbf{x}}' \widehat{\boldsymbol{\theta}}$. Also, from the Gauss-Markov theorem, an unbiased estimate of σ_ε^2 is provided by

$$\widehat{\sigma}_\varepsilon^2 = MSE. \quad (\text{A.196})$$

Therefore, a consistent estimate of ρ_{ε^2} is provided by

$$\hat{\rho}_{\varepsilon^2}^2 = R_{\varepsilon^2}^2 = \frac{\frac{1}{N} \sum_i (\mathbf{x}'_i \hat{\boldsymbol{\theta}} - \bar{y})^2}{\frac{1}{N} \sum_i (\mathbf{x}'_i \hat{\boldsymbol{\theta}} - \bar{y})^2 + MSE}. \quad (\text{A.197})$$

This is approximately equal to what is termed the adjusted R^2 in a multiple regression model, computed as

$$R_{adj}^2 = \frac{S_y^2 - MSE}{S_y^2}, \quad (\text{A.198})$$

where S_y^2 is the sample variance of Y .

When the estimated model does not explain any of the variation in Y , then $(\hat{\beta}_1, \dots, \hat{\beta}_p) = \mathbf{0}$ and $\mathbf{x}'_i \hat{\boldsymbol{\theta}} = \alpha = \bar{y}$ for all $i = 1, \dots, N$ and thus $\hat{\rho}_{\varepsilon^2}^2 = 0$. Conversely, when the estimated model explains all the variation in Y , then $\mathbf{x}'_i \hat{\boldsymbol{\theta}} = y_i$ for all observations and thus $\hat{\rho}_{\varepsilon^2}^2 = 1$.

A.8.2 Residual Variation

Another estimator of the explained loss using any function $\mathcal{L}(y_i, \tilde{y}_i)$ is the explained residual variation

$$R_{\mathcal{L}, resid}^2 = \frac{\sum_i \mathcal{L}(y_i, \hat{y}_0) - \sum_i \mathcal{L}(y_i, \hat{y}_i)}{\sum_i \mathcal{L}(y_i, \hat{y}_0)}, \quad (\text{A.199})$$

where \hat{y}_0 is the estimated unconditional prediction function free of any covariates, and \hat{y}_i the prediction function conditional on covariate vector \mathbf{x}_i , for which the loss function is minimized. For squared error loss with $\hat{y}_i = \hat{y}_i = \hat{E}(Y|\mathbf{x}_i)$ and $\hat{y}_0 = \bar{y}$, then

$$R_{\varepsilon^2, resid}^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS_y - SSE}{SS_y}. \quad (\text{A.200})$$

This is the traditional measure R^2 employed in multiple linear regression models.

From the partitioning of sums of squares (A.4), it follows that

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2, \quad (\text{A.201})$$

which yields

$$R_{\varepsilon^2, resid}^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}, \quad (\text{A.202})$$

and

$$E[R_{\varepsilon^2, resid}^2] = \rho_{\varepsilon^2}. \quad (\text{A.203})$$

In a multiple regression model $R_{\varepsilon^2, resid}^2$ is asymptotically equal to the $\hat{\rho}_{\varepsilon^2}^2$ presented in (A.197).

A.8.3 Negative Log-Likelihood Loss

Similar methods can be derived for special cases where a loss function other than squared error loss may have special meaning. One such case is the use of the entropy loss function in conjunction with the logistic regression model as described in Section 7.5.2. In this case, the expression for the fraction of explained residual entropy loss in a logistic regression model reduces to

$$R_\ell^2 = \frac{-2 \log L(\hat{\alpha}) - \left[-2 \log L(\hat{\alpha}, \hat{\beta}) \right]}{-2 \log L(\hat{\alpha})} = \frac{\text{Model } X_{LR}^2}{-2 \log L(\hat{\alpha})}. \quad (\text{A.204})$$

Although this measure was originally justified using entropy loss in logistic regression (Efron, 1978), this is a general measure of explained negative log likelihood that can be used with any likelihood-based regression model.

Consider a model with the conditional distribution $f(y; \mathbf{x}, \boldsymbol{\theta})$ as a function of a parameter vector $\boldsymbol{\theta}$. In logistic regression, for example, Y is a binary variable and $f(y; \mathbf{x}, \boldsymbol{\theta})$ is the Bernoulli distribution with probability π that is a function of the parameter vector $\boldsymbol{\theta} = (\alpha // \beta)$. The model negative log likelihood can then be expressed as

$$-\log L(\boldsymbol{\theta}) = \sum_i -\log f(y_i; \mathbf{x}, \boldsymbol{\theta}) = \sum_i \mathcal{L}(y_i, \tilde{y}_i) \quad (\text{A.205})$$

or as a sum of loss with the loss function $\mathcal{L}(y, \tilde{y}) = -\log f(y; \boldsymbol{\theta})$, where \tilde{y} is a function of $(\mathbf{x}, \boldsymbol{\theta})$.

Thus, the explained negative log likelihood can be used with any regression model where the corresponding loss function is $-\log f(y; \boldsymbol{\theta})$.

A.8.4 Madalla's R_{LR}^2

Another measure of explained variation initially proposed by Madalla (1983) and reviewed by Magee (1990) can be derived as a function of the likelihood ratio chi-square statistic. In the usual homoscedastic normal errors multiple linear regression model, as shown in the preceding example, the standard definition of R^2 is actually the $R_{\epsilon^2, \text{resid}}^2$ presented in (A.200). In this model, the null and full model log likelihoods are readily shown to be

$$\begin{aligned} \log [L(\hat{\alpha})] &= C - (N/2) \log (SS_y) \\ \log [L(\hat{\alpha}, \hat{\beta})] &= C - (N/2) \log (SSE), \end{aligned} \quad (\text{A.206})$$

where C is a constant that does not involve α or β . Then the p df model likelihood ratio test for the regression coefficients in this model is

$$\begin{aligned} X_{LR}^2 &= -2 \log \left[\frac{L(\hat{\alpha})}{L(\hat{\alpha}, \hat{\beta})} \right] = 2 \log L(\hat{\alpha}, \hat{\beta}) - 2 \log L(\hat{\alpha}) \\ &= N \log \left(\frac{SS_y}{SSE} \right) = N \log \left(\frac{1}{1 - R^2} \right) = -N \log (1 - R^2). \end{aligned} \quad (\text{A.207})$$

Therefore,

$$\exp[-X_{LR}^2/N] = 1 - R^2 \quad (\text{A.208})$$

and the standard measure of R^2 in normal errors models can also be derived as a *likelihood ratio* R^2 :

$$R_{LR}^2 = 1 - \exp[-X_{LR}^2/N]. \quad (\text{A.209})$$

This definition of R_{LR}^2 can also be applied to other models, such as logistic regression, where the likelihood ratio test is employed to test $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$, so that

$$\begin{aligned} R_{LR}^2 &= 1 - \exp\left(\frac{-X_{LR}^2}{N}\right) = 1 - \exp\left[\frac{2}{N} \log \frac{L(\theta_0)}{L(\hat{\theta})}\right] \\ &= 1 - \left[\frac{L(\theta_0)}{L(\hat{\theta})}\right]^{2/N} = R_M^2. \end{aligned} \quad (\text{A.210})$$

The latter expression is also known as Madalla's R_M^2 . This is a generally applicable measure of R^2 that can be applied to any model for a sample of N independent observations.

A.9 ROBUST INFERENCE

A.9.1 Information Sandwich

It is well established that least squares and maximum likelihood estimators are in general robust to specification of the variance structure and remain consistent even when the variance structure is misspecified (Huber, 1967). For example, in the ordinary homoscedastic errors least squares model we assume that the errors are *i.i.d.* with mean zero and common variance σ_ϵ^2 . If, instead, the errors have variance structure with $\text{Cov}(\epsilon) = \text{Cov}(y|\mathbf{x}) = \Sigma_\epsilon^2 \neq \sigma_\epsilon^2 \mathbf{I}_N$, then the ordinary least squares estimates of the coefficients in the model are still consistent and are still asymptotically normally distributed. The problem is that the correct expression for the variance of the coefficient estimates is the weighted least squares variance, not the OLS variance. That is, $V(\hat{\theta}) = (\mathbf{X}' \Sigma_\epsilon^{-1} \mathbf{X})^{-1} \neq (\mathbf{X}' \mathbf{X})^{-1} \sigma_\epsilon^2$. Thus, if the error variance is misspecified, then confidence limits and Wald tests that rely on a model-based estimate of the variance of the coefficients are distorted. In such instances it is preferable that the variances be estimated by a procedure that is robust to misspecification of the error variance structure.

Huber (1967), Kent (1982), and White (1982), among others, considered various aspects of robust maximum likelihood inference in which they explored properties of the maximum likelihood estimators when the likelihood is misspecified and suggested robust approximations to the likelihood ratio test. Based on these developments, Royall (1986) described the application of a simple robust estimate of the variance of the *MLEs* that can be used to protect against model misspecification. This estimate

has since come to be known as the *information sandwich* and is widely used in conjunction with such techniques as quasi-likelihood and generalized estimating equations (GEE) that are described subsequently. The information sandwich can also be used in conjunction with ordinary maximum likelihood estimation.

A.9.1.1 Correct Model Specification Consider a model in a scalar parameter θ where the model is correctly specified, meaning that the correct likelihood is specified. Then as shown in (A.113) of Section A.6.5, given the true value θ , the score $U(\theta) \stackrel{d}{\approx} \mathcal{N}[0, I(\theta)]$. From the Taylor's approximation of $U(\hat{\theta})$ about the true value θ in (A.114), it follows from (A.115) that asymptotically

$$\sqrt{n}(\hat{\theta} - \theta) \cong \frac{U(\theta)/\sqrt{n}}{-U'(\theta)/n} = \frac{\sum_i U_i(\theta)/\sqrt{n}}{-\sum_i U'_i(\theta)/n}. \quad (\text{A.211})$$

Rather than simplify as in (A.118), consider the limiting distribution of this ratio.

The total score $U(\theta)$ in the numerator is a sum of mean zero *i.i.d.* random variables with variance $V[U(\theta)] = \sum_i V[U_i(\theta)]$, where

$$V[U_i(\theta)] = E[U_i(\theta)^2] - \{E[U_i(\theta)]\}^2 = E[U_i(\theta)^2]. \quad (\text{A.212})$$

Then

$$V[U(\theta)] = nE[U_i(\theta)^2] = \sum_i E[U_i(\theta)^2] = E[i(\theta)] \quad (\text{A.213})$$

from (A.103). Thus, the numerator of (A.211) is asymptotically normally distributed with mean zero and variance

$$\lim_{n \rightarrow \infty} V \left[\sum_i \frac{U_i(\theta)}{\sqrt{n}} \right] = \lim_{n \rightarrow \infty} \frac{V[U(\theta)]}{n} = \lim_{n \rightarrow \infty} \frac{E[i(\theta)]}{n}. \quad (\text{A.214})$$

The denominator of (A.211) is the mean of the observed information for each observation in the sample where $i(\theta) = -\sum_i U'_i(\theta)$ is a sum of *i.i.d.* random variables. Thus, from the law of large numbers it follows that

$$\frac{i(\theta)}{n} \xrightarrow{p} E[-U'_i(\theta)] = \frac{I(\theta)}{n}. \quad (\text{A.215})$$

Using these results with Slutsky's convergence theorem (A.44), it follows that the *MLE* is asymptotically distributed as

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}[0, n\sigma_R^2(\theta)], \quad (\text{A.216})$$

with large sample variance

$$\sigma_R^2(\theta) = \frac{V[U(\theta)]}{I(\theta)^2} = \frac{\sum_i E[U_i(\theta)^2]}{I(\theta)^2} = \frac{E[i(\theta)]}{I(\theta)^2}, \quad (\text{A.217})$$

which is consistently estimated as

$$\hat{\sigma}_R^2(\hat{\theta}) = \frac{\sum_i [U_i(\hat{\theta})^2]}{I(\hat{\theta})^2}. \quad (\text{A.218})$$

From (A.213) the numerator is the empirical estimate of the observed information whereas the denominator is the square of the model-based estimate of the expected information.

The term *information sandwich* arises from the corresponding expressions in the vector parameter case. Following similar developments for a parameter vector with true value θ ,

$$(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}[0, \Sigma_R(\hat{\theta})], \quad (\text{A.219})$$

where the large sample variance-covariance matrix of the estimates is

$$\Sigma_R(\hat{\theta}) = I(\theta)^{-1} \left(\sum_i E[U_i(\theta)U_i(\theta)^T] \right) I(\theta)^{-1}, \quad (\text{A.220})$$

that is consistently estimated as

$$\hat{\Sigma}_R(\hat{\theta}) = I(\hat{\theta})^{-1} \left(\sum_i [U_i(\hat{\theta})U_i(\hat{\theta})^T] \right) I(\hat{\theta})^{-1}. \quad (\text{A.221})$$

The estimator is a "sandwich" where the bread is the model-based variance of the estimates and the meat is the empirical estimate of the observed information.

Thus, when the model is correctly specified, the score $U(\theta) \xrightarrow{d} \mathcal{N}[0, I(\theta)]$ where the variance can be consistently estimated either as $I(\hat{\theta})$ or by using the inverse information sandwich $\hat{\Sigma}_R(\hat{\theta})^{-1}$. Thus, the covariance matrix of the estimates can also be consistently estimated as either $I(\hat{\theta})^{-1}$, $I(\hat{\theta})^{-1}$, or $\hat{\Sigma}_R(\hat{\theta})$.

A.9.1.2 Incorrect Model Specification Now consider a "working model" in a parameter θ (possibly a vector) where the likelihood (or quasi-likelihood) used as the basis for developing or fitting the model is not correctly specified, so that it differs from the true likelihood in some respects. For example, the true likelihood may involve additional parameters, such as an overdispersion parameter, that is not incorporated into the working model likelihood. We assume, however, that the parameter of interest θ has the same meaning in both the working and true models. Kent (1982) then shows the following developments.

Given the true value θ , we can again apply a Taylor's expansion to yield (A.211). Even though the model is misspecified, $U(\theta) = \sum_i U_i(\theta)$ is still a sum of mean zero *i.i.d.* random variables, and thus is asymptotically normally distributed as

$$U(\theta) \xrightarrow{d} \mathcal{N}[0, J(\theta)], \quad (\text{A.222})$$

where $J(\theta) = nE[U_i(\theta)^2]$. Here the score is derived under the working model, but the expectation is with respect to the correct model. When the model is correctly specified, then $J(\theta) = I(\theta)$. However, when the working model is incorrect, then

$$J(\theta) = \sum_i V[U_i(\theta)] = \sum_i E[U_i(\theta)^2], \quad (\text{A.223})$$

since $E[U_i(\theta)] = 0$ under both the working and true models. Since the correct model is unknown, the actual expression for $J(\theta)$ is also unknown.

Likewise, the denominator in (A.211) involves the observed information computed using the first derivative of the scores under the working model. Nevertheless, this is the sum of *i.i.d.* random variables, and from the law of large numbers converges to

$$\lim_{n \rightarrow \infty} \sum_i -U'_i(\theta) \xrightarrow{P} K(\theta) = nE[-U'_i(\theta)], \quad (\text{A.224})$$

where the expectation is again taken with respect to the correct model. The expression for $K(\theta)$ is also unknown because the correct model is unknown. Then, from (A.211) it follows that asymptotically $\hat{\theta}$ is distributed as in (A.216) with large sample variance

$$\sigma_R^2(\theta) = \frac{J(\theta)}{K(\theta)^2}. \quad (\text{A.225})$$

For example, if the working model assumes that the distribution of the observations is $f(y; \theta)$ but the correct distribution is $g(y; \theta)$, then the likelihood function and score equations are derived from $f(\theta)$ but $J(\theta)$ and $K(\theta)$ are defined by the expectations of $U_i(\theta)^2$ and $U'_i(\theta)$ with respect to the correct distribution $g(\theta)$. Thus, for example, the term in $J(\theta)$ is of the form

$$E[U_i(\theta)^2] = \int_y \left(\frac{d \log f(y; \theta)}{d\theta} \right)^2 g(y; \theta) dy. \quad (\text{A.226})$$

In this case $J(\theta)$ and $K(\theta)$ are different from the expected information $I(\theta)$ under the working model. Also, note that $K(\theta)$ is the observed information with respect to $g(\theta)$.

Equivalent results are also obtained in the multiparameter case, yielding the matrices $\mathbf{J}(\theta)$ and $\mathbf{K}(\theta)$. Again, using the central limit theorem and Slutsky's convergence theorem, it follows that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}[0, n\mathbf{K}(\theta)^{-1}\mathbf{J}(\theta)\mathbf{K}(\theta)^{-1}]. \quad (\text{A.227})$$

Even though the expressions for $\mathbf{J}(\theta)$ and for $\mathbf{K}(\theta)$ are unknown, regardless of which model is correct, the variance can be consistently estimated using the empirical estimates

$$\begin{aligned} \hat{\mathbf{J}}(\theta) &= \sum_i [\mathbf{U}_i(\theta)\mathbf{U}_i(\theta)^T] \\ \hat{\mathbf{K}}(\theta) &= \sum_i -\mathbf{U}'_i(\theta). \end{aligned} \quad (\text{A.228})$$

In most cases, however, such as an overdispersed regression model or a quasi-likelihood regression model, it is assumed that the first moment specification is the same for the working and the correct models and that both models belong to the exponential family. In this case, Kent (1982) shows that $K(\theta) = \mathbf{I}(\theta)$ under the assumed model and

$$V(\hat{\theta}) = \Sigma_R(\hat{\theta}) = \mathbf{I}(\theta)^{-1}\mathbf{J}(\theta)\mathbf{I}(\theta)^{-1} \quad (\text{A.229})$$

in (A.220), which is consistently estimated using the information sandwich

$$\widehat{\Sigma}_R(\widehat{\theta}) = \mathbf{I}(\widehat{\theta})^{-1} \widehat{\mathbf{J}}(\widehat{\theta}) \mathbf{I}(\widehat{\theta})^{-1}, \quad (\text{A.230})$$

as in (A.221).

Thus, if the first moment model specification is correct but the second moment specification may be incorrect, the parameter estimates $\widehat{\theta}$ are still consistent estimates even if the working model error structure is incorrect. Further, the information sandwich provides a consistent estimate of the correct variance of the parameter estimates. Thus, tests of significance and confidence limits computed using the information sandwich variance estimates are asymptotically correct.

Example A.15 Poisson-Distributed Counts

For a sample of count data that is assumed to be distributed as Poisson, then from Example A.9 the score for each observation is

$$U_i(\theta) = \frac{x_i - \theta}{\theta}, \quad (\text{A.231})$$

and

$$\mathbf{I}(\theta)^{-1} = V(\widehat{\theta}) = \theta/N. \quad (\text{A.232})$$

Thus, the robust information sandwich estimator is

$$\widehat{\sigma}_R^2(\widehat{\theta}) = \frac{\sum_i (x_i - \widehat{\theta})^2}{N^2}. \quad (\text{A.233})$$

This estimator is consistent when the data are not distributed as Poisson, but rather have a variance that differs from the Poisson mean.

For the hospital mortality data in Example A.10, this estimator yields $\widehat{\sigma}_R^2(\widehat{\theta}) = 1.77$, which is about double the model-based estimate of 0.91, suggesting overdispersion in these data.

Example A.16 Homoscedastic Normal Errors Regression

Likewise, in ordinary multiple regression that assumes homoscedastic normally distributed errors, from Example A.11 the score for the i th observation with covariate vector \mathbf{x}_i , that includes the constant for the intercept, is

$$U_i(\theta) = \frac{\mathbf{x}_i(y_i - \mathbf{x}'_i \theta)}{\sigma_\epsilon^2}, \quad (\text{A.234})$$

and

$$\sum_i [U_i(\widehat{\theta}) U_i(\widehat{\theta})^T] = \frac{\mathbf{X}' \{ \text{diag}[(y_i - \mathbf{x}'_i \widehat{\theta})^2] \} \mathbf{X}}{(\sigma_\epsilon^2)^2} = \frac{\mathbf{X}' \widehat{\Sigma}_\epsilon \mathbf{X}}{(\sigma_\epsilon^2)^2}, \quad (\text{A.235})$$

where

$$\widehat{\Sigma}_\epsilon = \text{diag} \left[(y_i - \mathbf{x}'_i \widehat{\theta})^2 \right]. \quad (\text{A.236})$$

Given the model-based variance

$$V(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\boldsymbol{\theta})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} \sigma_e^2, \quad (\text{A.237})$$

then the robust information sandwich variance estimate is

$$\hat{\Sigma}_R(\hat{\boldsymbol{\theta}}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \hat{\Sigma}_e \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}, \quad (\text{A.238})$$

where the meat of the sandwich is the empirical estimate of the diagonal matrix of the variances of the errors.

This estimate then is consistent when the errors are not homoscedastic. This estimator was derived by White (1980) and is provided by SAS PROC REG using the ACOV specification. White (1980) also described a test of homoscedasticity that is also computed by SAS PROC REG using the *spec* option.

A.9.2 Robust Confidence Limits and Tests

The robust sandwich estimate of the variance of the *MLEs* can be used to provide large sample confidence limits that are robust to departures from the variance structure implied by the assumed model. This robust variance, also called the empirical variance, can also be used as the basis for robust Wald tests and robust efficient score tests. However, because the likelihood ratio test depends explicitly on the complete model specification, including the variance structure, a robust likelihood ratio test cannot be computed directly.

From (A.222), it follows that a robust score test of $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ for the parameter vector $\boldsymbol{\theta}$, or a $C(\alpha)$ -test for a subset of elements of $\boldsymbol{\theta}$, may be obtained as described in Section A.7.3 using the empirical estimate of the covariance matrix $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}_0)$ evaluated with respect to the parameter vector estimate $\hat{\boldsymbol{\theta}}_0$ obtained under the tested hypothesis. The properties of the robust score test have not been studied for logistic regression and Poisson regression. However, the robust score test and robust confidence limits in the Cox proportional hazards model were derived by Lin and Wei (1989) and Lin (1994); see Section 9.4.5. These computations are available as part of the SAS PROC PHREG; see Section 9.4.10.

A.10 GENERALIZED LINEAR MODELS AND QUASI-LIKELIHOOD

Generalized linear models (GLMs) refer to a family of regression models described by Nelder and Wedderburn (1972), which are based on distributions from the exponential family that may be fit using maximum likelihood estimation. This family of *GLMs* includes the normal errors, logistic, and Poisson regression models as special cases. For any distribution from the exponential family, the score equation and the observed and expected information are a function of the first and second moments only. This observation led to development of the method of quasi-likelihood by Wedderburn (1974) for fitting models that are not based on an explicit likelihood. Any model based on specification of the first two moments, including models that do not arise

from the exponential family, can be fit by the method of quasi-likelihood. The family of *GLMs* and quasi-likelihood estimation are elaborated in the text by McCullagh and Nelder (1989). An excellent general reference is Dobson (1990). Such models can be fit by the SAS PROC GENMOD and other programs.

A.10.1 Generalized Linear Models

In the simple normal errors linear regression model, the structural component is specified to be $y = \mathbf{x}'\boldsymbol{\theta} + \varepsilon$ or $y = \mu(\mathbf{x}) + \varepsilon$, where $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\theta}$. The random component of the model that describes the error distribution is then specified to be $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. This, in turn, specifies that the conditional distribution is $y|\mathbf{x} \sim \mathcal{N}(\mu(\mathbf{x}), \sigma_\varepsilon^2)$. A measure of the goodness of fit of the model is provided by the *deviance*, which is defined as the difference between the log likelihood of the present model and that of a model with as many parameters as necessary to fit the data perfectly. Thus, if there are N independent observations, the deviance is the difference between the log likelihood of a model with N *df* (thus fitting perfectly) and that of the present model with p *df*, the deviance having $N - p$ *df*. For a normal errors linear model the deviance equals the SSE/σ_ε^2 for that model.

The *GLM* family of models generalizes this basic model in two ways. First, the conditional distribution of $Y|\mathbf{x}$ can be any member of the exponential family. This includes the normal, binomial, Poisson, gamma, and so on. Second, the link between the conditional expectation μ and the linear predictor $(\mathbf{x}'\boldsymbol{\theta})$ can be any differentiable monotone function $g(\mu) = \mathbf{x}'\boldsymbol{\theta}$, called the link function, that maps the domain of μ onto the real line. Thus, $\mu(\mathbf{x}) = g^{-1}(\mathbf{x}'\boldsymbol{\theta})$, which we designate as $\mu_{\mathbf{x}}$. To allow for the variance to be an explicit function of the mean, we denote the error variance as $\sigma_\varepsilon^2(\mu_{\mathbf{x}})$.

Therefore, a *GLM* has three components:

1. The systematic component: e.g., $\eta = \mathbf{x}'\boldsymbol{\theta}$ = the linear predictor.
2. The link function $g(\mu) = \eta$ that specifies the form of the relationship between $E(Y|\mathbf{x}) = \mu_{\mathbf{x}}$ and the covariates \mathbf{x} .
3. The random component or the conditional distribution specification $y|\mathbf{x} \sim f(y; \varphi)$, where $f(\cdot)$ is a member of the exponential family.

The random component specification implies a specific relationship between the conditional mean $\mu_{\mathbf{x}}$ and the conditional variance $V(Y|\mathbf{x})$, that equals the variance of the errors, σ_ε^2 . The common choices are

1. *Normal* error distribution, which assumes that σ_ε^2 is constant for all \mathbf{x} and thus is statistically independent of $\mu_{\mathbf{x}}$ for all \mathbf{x} .
2. *Binomial* error distribution, which assumes that the error variance for the i th observation with "number of trials" n_i is of the form $\sigma_\varepsilon^2(\mu_{\mathbf{x}}) = n_i \mu_{\mathbf{x}} (1 - \mu_{\mathbf{x}})$, that is a function of $\mu_{\mathbf{x}}$. When $n_i = 1$ the distribution is Bernoulli.

3. *Poisson* error distribution, which assumes that $\sigma_\varepsilon^2(\mu_x) = \mu_x$.

In some cases it is also necessary to incorporate a scale or dispersion factor into the model, designated as ν . In addition, the *GLM* allows for observations to be weighted differentially as a function of a weight w_i for the i th observation, as where the i th observation is the mean of n_i measurements, in which case $w_i = n_i$.

A.10.2 Exponential Family of Models

The above specifications all fall within the framework of a regression model for the conditional expectation of a member of the exponential family of distributions that includes the normal, binomial, and Poisson as special cases. Thus, the estimating equations and estimated information for this family of models can be derived from the exponential family of distributions.

The probability function for the canonical form of the exponential family for the i th observation is

$$f(y_i; \varphi_i, \nu, w_i) = \exp \left[\frac{y_i \varphi_i - b(\varphi_i)}{a(\nu, w_i)} + c(y_i, \nu, w_i) \right], \quad (\text{A.239})$$

where the parameter of interest for the i th observation is φ_i , ν is a scale or dispersion parameter and w_i is a weighting constant. The functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ corresponding to the different members of the exponential family are readily derived and are presented in many basic texts on mathematical statistics. As we show below, the specific expressions are not needed for a family of regression models, the above *GLM* specifications being sufficient.

For a distribution of this form, the score for the i th subject is

$$U_i(\varphi_i) = \frac{d\ell_i}{d\varphi_i} = \frac{y_i - b'(\varphi_i)}{a(\nu, w_i)}, \quad (\text{A.240})$$

where $b'(\varphi) = db(\varphi)/d\varphi$. Likewise, the second derivative of the score for the i th observation is

$$U_i'(\varphi_i) = \frac{-b''(\varphi_i)}{a(\nu, w_i)}, \quad (\text{A.241})$$

where b'' refers to the second derivative. Since $E[U_i(\varphi_i)] = 0$, then (A.240) implies that

$$E(y_i) = \mu_i = b'(\varphi_i). \quad (\text{A.242})$$

To derive the expression for $V(y_i)$ note that Fisher's information equality (A.93) states that

$$E[U_i'(\varphi_i)] + E[U_i(\varphi_i)^2] = 0. \quad (\text{A.243})$$

Since $E[U_i(\varphi_i)] = 0$ for all observations, then $V(U_i(\varphi_i)) = E[U_i(\varphi_i)^2]$. Also, from (A.240),

$$V[U_i(\varphi_i)] = \frac{V(y_i)}{a(\nu, w_i)^2}. \quad (\text{A.244})$$

Therefore,

$$\frac{-b''(\varphi_i)}{a(\nu, w_i)} + \frac{V(y_i)}{a(\nu, w_i)^2} = 0, \quad (\text{A.245})$$

and thus

$$V(y_i) = b''(\varphi_i) a(\nu, w_i). \quad (\text{A.246})$$

Now consider that a *GLM* has been specified such that $E(Y|\mathbf{x}) = \mu_{\mathbf{x}}$ with link function

$$g(\mu_{\mathbf{x}}) = \eta_{\mathbf{x}} = \mathbf{x}'\boldsymbol{\theta} \quad (\text{A.247})$$

in terms of a covariate vector \mathbf{x} and coefficient vector $\boldsymbol{\theta}$. From the specified distribution of $Y|\mathbf{x}$, the conditional variance of $Y|\mathbf{x}$ equals the variance of the errors under that distribution and may be a function of the conditional expectation, or $V(Y|\mathbf{x}) = \sigma_{\varepsilon}^2(\mu_{\mathbf{x}})$. For the i th individual with covariate vector \mathbf{x}_i , we designate the conditional expectation as μ_i and variance as $V(y_i) = \sigma_{\varepsilon}^2(\mu_i)$. For a distribution from the exponential family, the score vector for the i th observation then is

$$\mathbf{U}_i(\boldsymbol{\theta}) = \frac{d\ell_i}{d\varphi_i} \frac{d\varphi_i}{d\mu_i} \frac{\partial\mu_i}{\partial\boldsymbol{\theta}}, \quad (\text{A.248})$$

where $\mu_i \equiv \mu(\mathbf{x}_i)$. From (A.240) and (A.242), the first term is

$$\frac{d\ell_i}{d\varphi_i} = \frac{y_i - E(y_i)}{a(\nu, w_i)} = \frac{y_i - \mu_i}{a(\nu, w_i)}. \quad (\text{A.249})$$

The second term is

$$\begin{aligned} \frac{d\varphi_i}{d\mu_i} &= \left(\frac{d\mu_i}{d\varphi_i} \right)^{-1} = \left(\frac{dE(y_i)}{d\varphi_i} \right)^{-1} = \left(\frac{db'(\varphi_i)}{d\varphi_i} \right)^{-1} \\ &= \frac{1}{b''(\varphi_i)} = \frac{a(\nu, w_i)}{\sigma_{\varepsilon}^2(\mu_i)}. \end{aligned} \quad (\text{A.250})$$

The final term is the vector

$$\frac{\partial\mu_i}{\partial\boldsymbol{\theta}} = \frac{d\mu_i}{d\eta_i} \frac{\partial\eta_i}{\partial\boldsymbol{\theta}} = \mathbf{x}_i \left(\frac{d\eta_i}{d\mu_i} \right)^{-1} = \frac{\mathbf{x}_i}{g'(\mu_i)}. \quad (\text{A.251})$$

Thus, the total score vector is

$$\mathbf{U}(\boldsymbol{\theta}) = \frac{\partial\ell}{\partial\boldsymbol{\theta}} = \sum_i \left[\frac{(y_i - \mu_i)}{\sigma_{\varepsilon}^2(\mu_i)} \right] \frac{\mathbf{x}_i}{g'(\mu_i)}, \quad (\text{A.252})$$

which provides the *MLE* estimating equation for $\boldsymbol{\theta}$.

The observed information is then obtained as

$$\mathbf{i}(\boldsymbol{\theta}) = -\mathbf{U}'(\boldsymbol{\theta}) = -\sum_i \frac{\partial}{\partial\beta} \left[\frac{d\ell_i}{d\varphi_i} \frac{d\varphi_i}{d\mu_i} \frac{\partial\mu_i}{\partial\boldsymbol{\theta}} \right]. \quad (\text{A.253})$$

After some algebra we obtain

$$\begin{aligned} i(\boldsymbol{\theta}) &= \sum_i \frac{\mathbf{x}_i \mathbf{x}'_i}{\sigma_\varepsilon^2(\mu_i) [g'(\mu_i)]^2} \\ &+ \sum_i (y_i - \mu_i) \mathbf{x}_i \mathbf{x}'_i \left(\frac{\sigma_\varepsilon^2(\mu_i) g''(\mu_i) + \left[\frac{d}{d\mu_i} \sigma_\varepsilon^2(\mu_i) \right] g'(\mu_i)}{[\sigma_\varepsilon^2(\mu_i)]^2 [g'(\mu_i)]^3} \right). \end{aligned} \quad (\text{A.254})$$

Since $E(y_i - \mu_i) = 0$, then the expected information is

$$\mathbf{I}(\boldsymbol{\theta}) = E[i(\boldsymbol{\theta})] = \sum_i \left[\frac{\mathbf{x}_i \mathbf{x}'_i}{\sigma_\varepsilon^2(\mu_i) g'(\mu_i)^2} \right]. \quad (\text{A.255})$$

This expression is also readily obtained as $E[\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T]$. Newton-Raphson iteration or Fisher scoring can then be used to solve for $\hat{\boldsymbol{\theta}}$, and to obtain the estimated observed and the estimated expected information. In some programs, such as SAS PROC GENMOD, the estimated covariance matrix of the coefficient estimates is computed using the inverse estimated observed information.

The above models include the ordinary normal errors multiple regression model, the logistic regression model, and the homogeneous Poisson regression model as special cases. In a two-parameter exponential distribution model, such as the homoscedastic normal errors model, φ is the mean and the scale parameter ν is the variance. In the binomial and Poisson models, the scale parameter is fixed at $\nu = 1$, although in these and other such models it is also possible to specify a scale or dispersion parameter $\nu \neq 1$ to allow for under or overdispersion. In such cases, however, we no longer have a "likelihood" in the usual sense. Nevertheless, such a model can be justified as a quasi-likelihood (see the next section).

In the binomial model, the weight may vary as a function of the number of "trials" (n_i) for the i th observation where $w_i = n_i$. Otherwise, the weight is usually a constant ($w_i = 1$). In these and other cases, $a(\nu, w_i) = \nu/w_i$.

In the exponential family, the expression for $\mu = b'(\varphi_i)$ also implies the *natural or canonical link* for that distribution such that $\varphi_i = g(\mu) = \eta_i = \mathbf{x}'_i \boldsymbol{\theta}$, thus leading to major simplifications in the above expressions. For the most common applications, the canonical link is as follows:

<i>Distribution</i>	$\mu = b'(\varphi)$	<i>Canonical Link, $g(\mu)$</i>
<i>Normal</i>	φ	<i>identity</i>
<i>Binomial</i>	$\frac{e^\varphi}{1 + e^\varphi}$	<i>logit</i>
<i>Poisson</i>	$\exp(\varphi)$	<i>log</i>

Although one could use any differentiable link function with any error distribution, problems may arise fitting the model by Newton-Raphson iteration. For example, one could use the log link with a binomial error distribution in lieu of the usual logistic regression model with a logit link. In this case, the elements of $\boldsymbol{\theta}$ describe

the covariate effects on the log risk or have an interpretation as log relative risks. However, this model does not ensure that the estimated probabilities $\pi(\mathbf{x}) = \mu(\mathbf{x})$ are bounded by $(0,1)$, and the iterative solution of the coefficients and the estimated information may fail unless a method for constrained optimization is used to fit the model.

A.10.3 Deviance and the Chi-Square Goodness of Fit

The model likelihood ratio test is constructed by comparing the $-2 \log [L(\alpha)]$ for the null intercept-only model to the $-2 \log [L(\alpha, \beta)]$ for the $(p+1)$ -variate model with parameter vector $\boldsymbol{\theta}$, where the difference is the likelihood ratio test statistic on p *df* under the model null hypothesis $H_0: \boldsymbol{\beta} = \mathbf{0}$. In *GLMs* the *deviance* is used to compare the $-2 \log [L(\alpha, \beta)]$ for the fit of the $(p+1)$ -variate model with that of a saturated model that fits the data perfectly on N *df*. The perfect-fit model, therefore, is one where $y_i = \mu_i$ for all $i = 1, \dots, N$. The deviance is the difference between these values on $(N - p - 1)$ *df*. Thus, the deviance provides a test of the hypothesis that the additional hypothetical $N - p - 1$ parameters are jointly zero, or that the model with parameters (α, β) is correct. In most cases, but not all, the deviance is also asymptotically distributed as chi-square under the hypothesis that the model fits the data.

For example, a normal errors linear model has log likelihood

$$\ell = \sum_{i=1}^N \left[\frac{-(y_i - \mu_i)^2}{2\sigma_\epsilon^2} - \log(\sqrt{2\pi}\sigma_\epsilon) \right], \quad (\text{A.256})$$

so that the perfect-fit model has log likelihood $\ell(\mu_1, \dots, \mu_N) = -N \log(\sqrt{2\pi}\sigma_\epsilon)$. Then, the deviance for a model where the conditional expectation $\mu(\mathbf{x}_i)$ is a function of a covariate vector \mathbf{x}_i reduces to

$$\begin{aligned} D(\alpha, \beta) &= -2\ell(\alpha, \beta) - [-2\ell(\mu_1, \dots, \mu_N)] \\ &= \sum_{i=1}^N \frac{[y_i - \mu(\mathbf{x}_i)]^2}{\sigma_\epsilon^2} = \frac{SS(\text{errors})}{\sigma_\epsilon^2}. \end{aligned} \quad (\text{A.257})$$

The deviance, therefore, is equivalent to the $-2 \log L(\alpha, \beta)$ less any constants. The greater the deviance, the greater the unexplained variation or lack of fit of the model.

In a logistic regression model where Y is a $(0,1)$ binary variable, the perfect-fit model log likelihood is

$$\ell(\mu_1, \dots, \mu_N) = \sum_{i=1}^N y_i \log \mu_i + \sum_{i=1}^N (1 - y_i) \log (1 - \mu_i) = 0, \quad (\text{A.258})$$

so that $D(\alpha, \beta) = -2\ell(\alpha, \beta)$. However, in a binomial regression model where y_i is also a count, the N *df* $\log L$ is not equal to zero. The same also applies to the

Poisson regression model where y_i is a count:

$$\begin{aligned}\ell(\mu_1, \dots, \mu_N) &= \sum_{i=1}^N [-\mu_i + y_i \log [\mu_i] - \log (y_i!)] \\ &= \sum_{i=1}^N [-y_i + y_i \log [y_i] - \log (y_i!)] \neq 0.\end{aligned}\quad (\text{A.259})$$

Likelihood ratio tests can also be computed as the difference between the deviances for nested models, the $-2\ell(\mu_1, \dots, \mu_N)$ canceling from both model deviances. Thus, even though the deviance itself may not be distributed as chi-square, the difference between deviances equals the likelihood ratio test that is asymptotically distributed as chi-square.

When the model is correctly specified and the deviance is asymptotically distributed as chi-square, then $E(\text{deviance}) = df = N - p - 1$. This provides a simple assessment of the adequacy of the second moment or mean-variance model specification. Thus, when the model variance assumptions apply, $E(\text{deviance}/df) = 1$. This also provides for a simple model adjustment to allow for overdispersion using a quasi-likelihood. However, the adequacy of this approach depends on the adequacy of the chi-square approximation to the large sample distribution of the deviance. For logistic regression models with a binary dependent variable, for example, this may not apply (see McCullagh and Nelder, 1989, pp. 118–119). Thus, in cases where the *deviance* is not approximately distributed as chi-square, the *deviance/df* should not be used as an indication of extra-variation.

Rather, it is preferred that the *Pearson chi-square test of goodness of fit* for the model be used to assess goodness of fit and the presence of extra-variation. For a *GLM* this statistic is the model test

$$X_P^2 = \sum_{i=1}^N \frac{[y_i - \hat{\mu}_i]^2}{\sigma_\varepsilon^2(\hat{\mu}_i)}, \quad (\text{A.260})$$

which is asymptotically distributed as chi-square on $N - p - 1$ *df* under the assumption that the model is correctly specified. Thus, for any model, $E(X_P^2) = df$ and X_P^2/df provides an estimate of the degree of over or underdispersion.

In general, it should be expected that the ratio *deviance/df* or X_P^2/df will vary by chance about 1 when the model is correctly specified. The variance of the chi-square statistic is $V(X^2) = 2(df)$, so that $V[X^2/df] = 2/df$. Thus, the range of variation expected when the model is correct with 95% confidence is on the order of $[1 \pm 1.96\sqrt{2/df}] = [1 \pm 2.77/\sqrt{df}]$, the approximate 95% tolerance limits. Thus, with $df = 100$ one should expect the ratio to fall within 1 ± 0.277 . One should then only consider adopting an overdispersed model when the ratio *deviance/df* or X_P^2/df departs substantially from 1.

A.10.4 Quasi-likelihood

In an important development, Wedderburn (1974) showed that the score equations in (A.252) could be used as what are termed *quasi-likelihood estimating equations*,

even when the precise form of the error distribution is unknown. All that is required for the asymptotic properties to apply is that the mean–variance relationship, or the first two moments of the conditional error distribution, be correctly specified. In this case it can be shown that the parameter estimates are asymptotically normally distributed about the true values with a covariance matrix equal to the inverse expected information, exactly as for maximum likelihood estimates (McCullagh, 1983, among others).

For example, assume that a set of quantitative observations is related to a vector of covariates \mathbf{X} with an identity link and the conditional error variance is constant for all values of \mathbf{X} , but the error distribution and the conditional distribution of $Y|\mathbf{x}$ are not a normal distribution. Even though the error distribution is not the normal distribution, the quasi-likelihood estimates obtained from a normal errors assumption are asymptotically normally distributed as in (A.119). This is not surprising since the assumptions (excluding normality) are the same as those required for fitting the linear model by ordinary least squares as described in Section A.5.1, which, in turn, are sufficient to show that the estimates are asymptotically normally distributed using the central limit theorem.

As a special case of a quasi-likelihood model, consider a *GLM* where the first moment for the i th observation with covariate vector \mathbf{x}_i is specified to be of the form $E(y_i|\mathbf{x}_i) = \mu_i$, where $g(\mu_i) = \eta_i = \mathbf{x}'_i \boldsymbol{\theta}$ and where the second moment is specified to be of the form $V(y_i|\mathbf{x}_i) = \nu\sigma^2(\mu_i) = \nu\sigma_i^2$ that can be some function of the conditional expectation with scale or dispersion parameter ν . We also allow the i th observation to have weight w_i . Then the quasi-likelihood estimate can also be derived as a *minimum chi-square estimate* of the parameters.

The Pearson chi-square statistic for the goodness of fit of the model, assuming known parameters, is

$$X^2 = \sum_{i=1}^N \frac{w_i(y_i - \mu_i)^2}{\nu\sigma_i^2}. \quad (\text{A.261})$$

If we ignore the possible dependence of the conditional variance on the conditional expectations, that is, we treat the $\{\sigma_i^2\}$ as fixed constants, then the minimum chi-square estimates are obtained as the solution to the estimating equation

$$\frac{\partial X^2}{\partial \boldsymbol{\theta}} = \sum_{i=1}^N \frac{2w_i(y_i - \mu_i)}{\nu\sigma_i^2} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\theta}} \right) = \mathbf{0}. \quad (\text{A.262})$$

This estimating equation is of the form $\tilde{\mathbf{U}}(\boldsymbol{\theta}) = \mathbf{0}$ in terms of a *quasi-score function* $\tilde{\mathbf{U}}(\boldsymbol{\theta})$, which can be expressed in matrix terms as

$$\tilde{\mathbf{U}}(\boldsymbol{\theta}) = \mathbf{D}' \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\mu}), \quad (\text{A.263})$$

where $\mathbf{D} = (\partial\mu_1/\partial\boldsymbol{\theta} \cdots \partial\mu_N/\partial\boldsymbol{\theta})^T$, $\mathbf{V} = \nu[\text{diag}(\sigma_1^2 \cdots \sigma_N^2)]$, $\mathbf{Y} = (y_1 \cdots y_N)^T$, and $\boldsymbol{\mu} = (\mu_1 \cdots \mu_N)^T$.

The family of models considered here is much broader than that in a *GLM* based on an error distribution from the exponential family. Further, in the event that the quasi-likelihood equals an exponential family likelihood, the quasi-likelihood score

equation $\tilde{U}(\boldsymbol{\theta})$ equals the total score from the exponential family likelihood presented in (A.252), with the simplification that $a(\nu, w_i) = \nu/w_i$.

It also follows that an estimate that minimizes the Pearson chi-square is also a weighted least squares estimate because the chi-square objective function X^2 is equivalent to the weighted sum of squares of errors that is minimized using weighted least squares. Thus, quasi-likelihood estimates are usually obtained using iteratively reweighted least squares (IRLS). Algebraically, it can also be shown that the systems of equations solved using IRLS are equivalent to those solved using Fisher scoring iteration (see Hillis and Davis, 1994).

Assuming that the model specifications are correct, it is then readily shown that

$$\begin{aligned} E[\tilde{U}(\boldsymbol{\theta})] &= \mathbf{0} \\ E[-\tilde{U}'(\boldsymbol{\theta})] &= \mathbf{D}'\mathbf{V}^{-1}\mathbf{D} \\ Cov[\tilde{U}(\boldsymbol{\theta})] &= \tilde{I}(\boldsymbol{\theta}) = \mathbf{D}'\mathbf{V}^{-1}\mathbf{D}. \end{aligned} \quad (\text{A.264})$$

Since the quasi-score function is a sum of *i.i.d.* random variables, then using the same developments as in Section A.6.5, it follows that the quasi score converges in distribution to the normal

$$\sqrt{n}\tilde{U}(\boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, n\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}], \quad (\text{A.265})$$

and the maximum quasi-likelihood estimates are asymptotically normally distributed as

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}]. \quad (\text{A.266})$$

From this asymptotic normal distribution of the estimates, it follows that Wald tests and confidence limits of the parameter estimates are readily obtained. In addition, quasi-score statistics are also readily constructed. Even though we do not start from a full likelihood, a quasi-likelihood ratio test can be constructed using the change in $-2 \log$ quasi-likelihood as for a full likelihood.

One advantage of quasi-likelihood estimation is that extra-variation or overdispersion can readily be incorporated into the model by assuming that a common variance inflation factor ν applies to all observations such that $V(y_i|\mathbf{x}_i) = \nu\sigma^2(\mu_i) = \nu\sigma_i^2$. As described in Section A.10.3, if we first fit a homogeneous (not overdispersed) model with the scale factor fixed at $\nu = 1$, then $E(X_P^2/df) = \nu$, where X_P^2 is the Pearson chi-square from the homogeneous model. Thus, a moment estimator of the variance inflation or overdispersion factor is $\hat{\nu} = X_P^2/df$. If this estimate is approximately equal to 1 in a logistic or Poisson regression model, then this indicates that the original model specifications apply. Otherwise, if $\hat{\nu}$ departs substantially from 1, such as outside the 95% tolerance limits $[1 \pm 2.77/\sqrt{df}]$, this is an indication that an under- or overdispersed model may be appropriate. In this case, a model can be refit based on the quasi-likelihood where ν is not fixed. Extra-variation, however, may also arise from a fundamental model misspecification or the omission of important covariates.

Although it is possible to solve a set of quasi-likelihood estimating equations jointly for $(\nu, \boldsymbol{\theta})$, this is not the approach generally employed. Rather, computer

programs such as the SAS PROC GENMOD use the moment estimating equation for ν to compute an iterative estimate of the model.

Finally, we note that the robust variance estimate may also be employed in conjunction with a quasi-likelihood model to obtain estimates of the covariance matrix of the parameter estimates that are robust to misspecification of the first and second moments of the quasi-likelihood. In this case,

$$\widehat{V}(\widehat{\theta}) = \tilde{I}(\widehat{\theta})^{-1} \widehat{J}(\widehat{\theta}) \tilde{I}(\widehat{\theta})^{-1} = \widehat{\Sigma}_R(\widehat{\theta}), \quad (\text{A.267})$$

where

$$\widehat{J}(\widehat{\theta}) = \sum_i [\tilde{U}(\widehat{\theta}) \tilde{U}(\widehat{\theta})^T]. \quad (\text{A.268})$$

This provides for robust confidence intervals and Wald tests, and also for a robust quasi-score test.

A.10.5 Conditional GLMs

Chapter 7 describes the conditional logistic regression model for matched sets with binary responses, and Chapter 8 that for the conditional Poisson regression model for count responses. Both are members of the family of *conditional* generalized linear models based on members of the exponential family. These models are also discussed in the general text by McCullagh and Nelder (1989). Another special case is the linear model for quantitative responses that are normally distributed within matched sets. Since the sufficient statistic for the matched set intercept is the sum of the observations within the set, the conditional normal likelihood is readily derived, from which estimating equations and inferences for the covariate parameters are readily obtained. Unfortunately, software is not available to fit this family of conditional *GLMs* for matched sets.

A.11 GENERALIZED ESTIMATING EQUATIONS (GEE)

One of the most important developments in recent years is the development of models based on *generalized estimating equations (GEEs)* for the analysis of correlated observations. Liang and Zeger (1986) and Zeger and Liang (1986) proposed fitting *GLM*-like models for correlated observations using a generalization of quasi-likelihood. The use of *GEEs* for the analysis of longitudinal data is reviewed in the text by Diggle et al. (1994), among others. This approach generalizes many of the methods presented in this text to the analysis of longitudinal data with repeated measurements, such as a longitudinal logistic regression model.

Assume that up to K repeated measurements at observation times t_1, \dots, t_K are obtained for each of N subjects. To allow for randomly missing observations, let $\mathbf{y}_i = (y_{i1} \ \dots \ y_{in_i})^T$ denote the vector of $n_i \leq K$ observed responses for the i th subject, $i = 1, \dots, N$. For each measurement there is an associated covariate vector \mathbf{x}_{ij} consisting of the constant plus p covariate values, that in turn may differ for each

repeated measure. Then let \mathbf{X}_i denote the $n_i \times (p+1)$ matrix of covariate values with row \mathbf{x}_{ij}^T for the j th replicate measure. The matrix \mathbf{X}_i may include time-dependent covariates under the usual assumption that the covariate values reflect the past history of the covariate process at time t_j (see Section 9.4.3).

The number of observed replicates may vary among subjects due to a variety of mechanisms for missing data. The simplest is administratively missing data, as in a study with staggered patient entry. For example, in a study with five annual visits where subjects are recruited over a two-year period, the first subject entered can have all five visits whereas the last can have only the first three visits, the last two being administratively missing. If subjects are sampled from a homogeneous population, then the missing data qualify as being missing completely at random (MCAR), meaning that the missing data do not depend either on what has been observed or what is missing (i.e., the hypothetical measurable value). Other mechanisms may also qualify as MCAR, although this may be unprovable. Alternatively, if it can be claimed that the failure to observe a response value is a function of other known and measured covariates, then the missing data can be claimed to be missing at random (MAR). Under these assumptions, *GEE* provides a valid analysis.

As in the family of *GLMs*, assume that $E(y_{ij}) = \mu_{ij}$, where $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\theta}$ and that $V(y_{ij}|\mathbf{x}_{ij}) = \nu\sigma^2(\mu_{ij})$ for each repeated measure marginally, $j = 1, \dots, n_i$. For the i th subject, let $\mathbf{A}_i = \text{diag}[\nu\sigma^2(\mu_{i1}) \dots \nu\sigma^2(\mu_{in_i})]$ denote the matrix of marginal variances. Then $\Sigma_i = \text{Cov}(\mathbf{y}_i)$ can be expressed as a function of \mathbf{A}_i and the correlation matrix among the repeated measures, say $\boldsymbol{\rho}$. Thus,

$$\Sigma_i = \mathbf{A}_i^{1/2} \boldsymbol{\rho} \mathbf{A}_i^{1/2}. \quad (\text{A.269})$$

The form of the correlation matrix, however, is unknown in practice. Thus, we adopt a working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ as a function of other parameters $\boldsymbol{\alpha}$ that are distinct from, and not to be confused with, the intercept component of $\boldsymbol{\theta}$. This yields a working covariance matrix

$$\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}. \quad (\text{A.270})$$

Generalizing the quasi-likelihood score equation in (A.263), the generalized estimating equation (*GEE*) for the coefficient vector $\boldsymbol{\theta}$ is

$$\tilde{\mathbf{U}}_G(\boldsymbol{\theta}) = \sum_{i=1}^N \mathbf{D}'_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \quad (\text{A.271})$$

where $\boldsymbol{\mu}_i = (\mu_{i1} \dots \mu_{in_i})^T$ and $\mathbf{D}_i = (\partial \boldsymbol{\mu}_i / \partial \boldsymbol{\theta})$.

There are many possible specifications for the working correlation structure. The most commonly employed are

1. **Independence.** Under the assumption of independence, $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}$ and each repeated measure is assumed independent of the others. With small sample sizes this structure is probably inappropriate if in fact there is some correlation among the replicates, and will result in a loss of efficiency in the coefficient estimates. However, with large samples there is little, if any, loss of efficiency.

2. **Exchangeable.** This model assumes that the correlation between all $K(K - 1)/2$ pairs of measures are a constant α , or that $\mathbf{R}(\alpha) = \mathbf{I} + \alpha \mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ is a ones vector. Note that the marginal variances, or the elements of \mathbf{A}_i , may vary among measures, so this assumption specifies that the covariances also vary among pairs of measures, but with a constant correlation. This is similar to the compound symmetry covariance structure as used in a repeated measures analysis of variance. However, compound symmetry is a more restrictive assumption because it specifies that the marginal variance is also constant over all measures and there is a constant covariance among all pairs of measures, resulting in an exchangeable correlation matrix.
3. **First-order autoregressive.** This AR(1) model assumes that $\mathbf{R}(\alpha) = \alpha^{|j-k|}$ for $j \neq k$, so that the correlation decreases as the "distance" between measurement points increases. This model may naturally apply to measurements taken over time since consecutive repeated measurements may be more strongly correlated than those measured over a longer period of time. However, the correlations decay quickly as the number of steps between measures increases. For example, if $\alpha = 0.5$ for one step, then the correlation is 0.25 at two steps, 0.125 at three steps, and so on.
4. **Unstructured.** For K measurements among subjects, the correlation matrix consists of $K(K - 1)/2$ elements. If this number is small relative to N , then the entire correlation matrix ρ can be estimated empirically, in which case $\mathbf{R}(\alpha) = \hat{\rho}$ and α is the set of correlation values.

From the theory of estimating functions (cf. Liang and Zeger, 1995), solution of the *GEE* estimating equations provides a consistent estimate of θ even though the covariance structure \mathbf{V}_i may be misspecified, either due to misspecification of the marginal variance structure \mathbf{A}_i or the correlation structure $\mathbf{R}(\alpha)$. However, the model-based estimated covariance matrix of the estimates, a generalization of that in (A.266), is not consistent. Thus, a generalization of the robust variance estimate is employed.

If the specified covariance structure \mathbf{V}_i is correct, then the model-based information function is obtained as a generalization of (A.264),

$$\mathbf{I}(\theta) = \sum_{i=1}^N \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i, \quad (\text{A.272})$$

that yields the model-based covariance matrix of the estimates $\hat{\theta}$ from the inverse information matrix. The empirical estimate of the covariance matrix of the estimates is provided by

$$\begin{aligned} \mathbf{J}(\theta) &= \sum_{i=1}^N \tilde{\mathbf{U}}_G(\theta) [\tilde{\mathbf{U}}_G(\theta)]^T = \sum_{i=1}^N \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) (\mathbf{y}_i - \boldsymbol{\mu}_i)' \mathbf{V}_i^{-1} \mathbf{D}_i \\ &= \sum_{i=1}^N \mathbf{D}_i' \mathbf{V}_i^{-1} \text{Cov}(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i. \end{aligned} \quad (\text{A.273})$$

Then the robust empirical covariance matrix of the estimates is provided by

$$\Sigma_R(\hat{\theta}) = \mathbf{I}(\theta)^{-1} \mathbf{J}(\theta) \mathbf{I}(\theta)^{-1}, \quad (\text{A.274})$$

that provides the basis for large sample empirical Wald tests of coefficient estimates and their confidence limits.

The model is fit recursively. Using the current estimates of the correlation parameters α and the scale parameter ν , the *GEE* estimating equations are solved iteratively to obtain an updated estimate $\hat{\theta}$. This is used to provide updated estimates of the residuals which then provide updated estimates of α and ν . Upon convergence, the equations in (A.272)–(A.274) are evaluated using $(\hat{\alpha}, \hat{\nu}, \hat{\theta})$ to provide the robust estimate $\hat{\Sigma}_R(\hat{\theta})$.

Rotnitsky and Jewel (1990) describe the properties of Wald and score tests in *GEE*, and Boos (1992) provides a generalization of efficient scores tests to a wide family of models, including *GEE*. Assume that we wish to test a hypothesis of the form $H_0: \mathbf{L}'\theta = \mathbf{0}$, where \mathbf{L} is a $(p+1) \times r$ matrix of r contrasts and where $\theta = (\alpha \ \beta)^T$ of $p+1$ parameters. To test the model, or the hypothesis that $\beta = \mathbf{0}$, then $\mathbf{L}' = (\mathbf{1}_p \parallel \mathbf{I}_p)$ where $r = p$, $\mathbf{1}_p$ is a column vector of ones, and \mathbf{I}_p is the p -dimension identity matrix. Also, let $\tilde{\mathbf{U}}_G(\hat{\theta}_0)$ denote the score vector evaluated using the parameter estimates obtained under the tested hypothesis. Then the value of the score test is provided by

$$X_{GS}^2 = \tilde{\mathbf{U}}_G(\hat{\theta}_0)' \mathbf{I}(\hat{\theta})^{-1} \mathbf{L} \left[\mathbf{L}' \Sigma_R(\hat{\theta}) \mathbf{L} \right]^{-1} \mathbf{L}' \mathbf{I}(\hat{\theta})^{-1} \tilde{\mathbf{U}}_G(\hat{\theta}_0), \quad (\text{A.275})$$

that is asymptotically distributed as χ_r^2 on r df.

References

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.*, 6, 701–726.
- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In, *Second International Symposium on Information Theory*. Petrov, B.N. and Csaki, F. (eds.), 267–281. Budapest: Akademiai Kiado.
- Alioum, A. and Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, 52, 512–524.
- Altschuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Math. Biosci.*, 6, 1–11.
- Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10, 1100–1120.
- Andersen, P.K. and Rasmussen, N.K. (1986). Psychiatric admissions and choice of abortion. *Statist. Med.*, 5, 243–253.
- Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1982). Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. *Int. Statist. Rev.*, 50, 219–258.
- Andersen, P.K., Borgan O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.
- Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika* 59, 19–35.

- Anderson, J.R. and Bernstein, L. (1985). Asymptotically efficient two-step estimators of the hazards ratio for follow-up studies and survival data. *Biometrics*, 41, 733–739.
- Anderson, T.W. (1984). *An Introduction to Multivariate Analysis*, 2nd ed. New York: John Wiley & Sons.
- Anscombe, F.J. (1956). On estimating binomial response relations. *Biometrika*, 43, 461–464.
- Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, 68, 357–363.
- Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* 11, 375–386.
- Bailar, J.C., Louis, T.A., Lavori, P.W. and Polansky, M. (1984). A classification for biomedical research reports. *N. Engl. J. Med.*, 311, 1482–1487.
- Bancroft, T.A. (1972). Some recent advances in inference procedures using preliminary tests of significance. In *Statistical Papers in Honor of George W. Snedecor*, Bancroft, T.A. (ed.), 19–30. Ames, IA: The Iowa State University Press.
- Barnard, G.A. (1945). A new test for 2×2 tables. *Nature*, 156, 177.
- Barnard, G.A. (1949). Statistical inference. *J. Roy. Statist. Soc., B*, 11, 115–139.
- Bartlett, M.S. (1955). Approximate confidence intervals: III. A bias correction. *Biometrika*, 42, 201–204.
- Beach, M.L. and Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Control. Clin. Trials*, 10, 161S–175S.
- Bean, S.J. and Tsokos, C.P. (1980). Developments in non-parametric density estimation. *Int. Statist. Rev.*, 48, 267–287.
- Bennett, S. (1983). Log-logistic regression models for survival data. *Appl. Statist.*, 32, 165–171.
- Bhapkar, V.P. (1979). On tests of marginal symmetry and quasi-symmetry in two- and three-dimensional contingency tables. *Biometrics*, 35, 417–426.
- Bickel, P. and Doksum, K. (1977). *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice-Hall.
- Birch, M.W. (1964). The detection of partial association: I. The 2×2 case. *J. Roy. Statist. Soc. B*, 26, 313–324.
- Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Blackwelder, W.C. (1982). Proving the null hypothesis in clinical trials. *Control. Clin. Trials*, 3, 345–353.
- Bloch, D.A. and Kraemer, H.C. (1989). 2×2 kappa coefficients: measures of agreement or association. *Biometrics*, 45, 269–287.
- Bloch, D.A. and Moses, L.E. (1988). Nonoptimally weighted least squares. *Amer. Statist.*, 42, 50–53.
- Blum, A.L. (1982). Principles for selection and exclusion. In *The Randomized Clinical Trial and Therapeutic Decisions*, Tygstrup, N., Lachin, J.M. and Juhl, E. (eds.), 43–58. New York: Marcel Dekker.

- Boos, D. (1992). On generalized score tests. *Am. Statistician*, 46, 327–333.
- Bowker, A.H. (1948). A test for symmetry in contingency tables. *J. Amer. Statist. Assoc.*, 43, 572–574.
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.*, 87, 738–754.
- Breslow, N.E. (1972). Discussion of Professor Cox's paper. *J. Roy. Statist. Soc. B*, 34, 216–217.
- Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89–99.
- Breslow, N.E. (1975). Analysis of survival data under the proportional hazards model. *Int. Statist. Rev.*, 43, 45–58.
- Breslow, N.E. (1981). Odds ratio estimators when the data are sparse. *Biometrika*, 68, 73–84.
- Breslow, N.E. (1982). Covariance adjustment of relative-risk estimates in matched studies. *Biometrics*, 38, 661–672.
- Breslow, N.E. (1984). Extra-Poisson variation in log-linear models. *Appl. Statist.*, 33, 38–44.
- Breslow, N.E. (1996). Statistics in epidemiology: the case-control study. *J. Amer. Statist. Assoc.*, 91, 14–28.
- Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research, Vol. 1, The Analysis of Case-Control Studies*. Oxford, UK: Oxford University Press.
- Breslow, N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research, Vol. 2, The Design and Analysis of Cohort Studies*. Oxford, UK: Oxford University Press.
- Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L. and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epidemiol.*, 108, 299–307.
- Bronowski, J. (1973). *The Ascent of Man*. Boston: Little Brown.
- Bross, I.D.J. (1958). How to use ridit analysis. *Biometrics*, 14, 18–38.
- Byar, D.P. (1985). Prognostic variables for survival in a randomized comparison of treatments for prostatic cancer. In *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Herzberg, A.M. and Andrews, D.F. (eds.), 261–274. New York: Springer-Verlag.
- Canner, P.L. (1991). Covariate adjustment of treatment effects in clinical trials. *Control. Clin. Trials*, 12, 359–366.
- CDP (Coronary Drug Project Research Group) (1974). Factors influencing long-term prognosis after recovery from myocardial infarction - Three-year findings of the Coronary Drug Project. *J. Chronic Dis.*, 27, 267–285.
- Chavers, B.M., Bilous, R.W., Ellis, E.N., Steffes, M.W. and Mauer, S.M. (1989). Glomerular lesions and urinary albumin excretion in type I diabetes without overt proteinuria. *N. Engl. J. Med.*, 319, 966–970.

- Cicchetti, D.V. and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *Am. J. EEG Technol.*, 11, 101–109.
- Clayton, D.G. (1974). Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika*, 61, 525–531.
- Clayton, D.G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *J. Roy. Statist. Soc. B*, 148, 82–117.
- Clopper, C.J. and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404–413.
- Cochran, W.G. (1950). The comparison of percentages in the matched samples. *Biometrika*, 37, 256–266.
- Cochran, W.G. (1954a). Some methods for strengthening the χ^2 tests. *Biometrics*, 10, 417–451.
- Cochran, W.G. (1954b). The combination of estimates from different experiments. *Biometrics*, 10, 101–129.
- Cochran, W.G. (1983). *Planning and Analysis of Observational Studies*. New York: John Wiley & Sons.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Ed. Psychol. Meas.*, 20, 37–46.
- Cohen, J. (1968). Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, 70, 213–220.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Laurence Erlbaum Associates.
- Collett, D. (1991). *Modelling Binary Data*. London: Chapman&Hall.
- Collett, D. (1994). *Modelling Survival Data in Medical Research*. London: Chapman&Hall.
- Collins, R., Yusuf, S. and Peto, R. (1985). Overview of randomised trials of diuretics in pregnancy. *Br. Med. J.*, 290, 17–23.
- Connett, J.E., Smith, J.A. and McHugh, R.B. (1987). Sample size and power for pair-matched case-control studies. *Statist. Med.*, 6, 53–59.
- Connor, R.J. (1987). Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43, 207–211.
- Conover, W.J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables. (with comments). *J. Amer. Statist. Assoc.*, 69, 374–382.
- Cook, R.J. and Sackett, D.L. (1995). The number needed to treat: a clinically useful measure of a treatment effect. *Br. Med. J.*, 310, 452–454.
- Cornfield, J. (1951). A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast, and cervix. *J. Natl. Cancer Inst.*, 11, 1269–1275.
- Cornfield, J. (1954). The estimation of the probability of developing a disease in the presence of competing risks. *Am. J. Public Health*, 47, 601–607.

- Cornfield, J. (1956). A statistical problem arising from retrospective studies. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4*, Neyman, J. (ed.), 135–148. Berkeley, CA: University of California Press.
- Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Fed. Proc.*, 21, 58–61.
- Cox, D.R. (1958a). The regression analysis of binary sequences. *J. Roy. Statist. Soc. B*, 20, 215–242.
- Cox, D.R. (1958b). Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.
- Cox, D.R. (1970). *The Analysis of Binary Data*. (2nd ed., 1989. Cox, D.R. and Snell, E. J.). London: Chapman&Hall.
- Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B*, 34, 187–220.
- Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman&Hall.
- Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman&Hall.
- Cox, D.R. and Miller, H.D. (1965). *The Theory of Stochastic Processes*. London: Chapman&Hall.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.
- Cutler, S.J. and Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *J. Chronic Dis.*, 8, 699–712.
- Davies, M. and Fleiss, J.L. (1982). Measuring agreement for multinomial data. *Biometrics*, 38, 1047–1051.
- Day, N.E. and Byar, D.P. (1979). Testing hypotheses in case-control studies: equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics*, 35, 623–630.
- Dean, C. and Lawless, J.F. (1989). Test for detecting over-dispersion in Poisson regression models. *J. Amer. Statist. Assoc.*, 84, 467–472.
- Deckert, T., Poulsen, J.E. and Larsen, M. (1978). Prognosis of diabetics with diabetes onset before the age of thirty-one: I. Survival, causes of death, and complications. *Diabetologia*, 14, 363–370.
- Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Hoboken, NJ: John Wiley & Sons.
- DerSimonian, R. and Laird, N.M. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials*, 7, 177–188.
- Desu, M.M. and Raghavarao, D. (1990). *Sample Size Methodology*. San Diego, CA: Academic Press.

- Diabetes Control and Complications Trial Research Group (DCCT) (1990). The Diabetes Control and Complications Trial (DCCT): update. *Diabetes Care*, 13, 427–433.
- Diabetes Control and Complications Trial Research Group (DCCT) (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.*, 329, 977–986.
- Diabetes Control and Complications Research Group (DCCT) (1995a). Effect of intensive therapy on the development and progression of diabetic nephropathy in the Diabetes Control and Complications Trial. *Kidney Int.*, 47, 1703–1720.
- Diabetes Control and Complications Trial Research Group (DCCT) (1995b). Adverse events and their association with treatment regimens in the Diabetes Control and Complications Trial. *Diabetes Care*, 18, 1415–1427.
- Diabetes Control and Complications Trial Research Group (DCCT) (1995c). The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the Diabetes Control and Complications Trial. *Diabetes*, 44, 968–983.
- Diabetes Control and Complications Trial Research Group (DCCT) (1995d). Progression of retinopathy with intensive versus conventional treatment in the Diabetes Control and Complication Trial. *Ophthalmology*, 102, 647–661.
- Diabetes Control and Complications Trial Research Group (DCCT) (1996). The absence of a glycemic threshold for the development of long-term complications: the perspective of the Diabetes Control and Complications Trial. *Diabetes*, 45, 1289–1298.
- Diabetes Control and Complications Trial Research Group (DCCT) (1997). Hypoglycemia in the Diabetes Control and Complications Trial. *Diabetes*, 46, 271–286.
- Diabetes Control and Complications Trial Research Group (DCCT) (2000). The effect of pregnancy on microvascular complications in the diabetes control and complications trial. *Diabetes Care*, 23, 1084–1091.
- Diabetes Prevention Program Research Group (DPP) (2002). Hypertension, insulin, proinsulin in participants with impaired glucose tolerance. *Hypertension*, 40, 679–686.
- Dick, T.D.S. and Stone, M.C. (1973). Prevalence of three cardinal risk factors in a random sample of men and in patients with ischaemic heart disease. *Br. Heart J.*, 35, 381–385.
- Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.
- Dobson, A. (1990). *An Introduction to Generalized Linear Models*. London: Chapman&Hall.
- Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials: a review. *Statist. Med.*, 3, 199–214.

- Dorn, H.F. (1944). Illness from cancer in the United States. *Public Health Rep.*, 59, Nos. 2, 3, and 4.
- Dyke, G.V. and Patterson, H.D. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics*, 8, 1–12.
- Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*, 351, 1451–1467.
- Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.*, 72, 557–565.
- Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. *J. Amer. Statist. Assoc.*, 73, 113–121.
- Efron, B. (1998). Fisher in the 21st century (with discussion). *Statist. Sci.*, 13, 95–122.
- Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimates: observed versus expected Fisher information. *Biometrika*, 65, 457–487.
- Efroymson, M.A. (1960). Multiple regression analysis. In *Mathematical Methods for Digital Computers*, Ralston, A. and Wilf, H.S. (eds.), 191–203. New York: John Wiley & Sons.
- Ejigou, A. and McHugh, R.B. (1977). Estimation of relative risk from matched pairs in epidemiological research. *Biometrics*, 33, 552–556.
- Elandt-Johnson, R.C. and Johnson, N.L. (1980). *Survival Models and Data Analysis*. New York: John Wiley & Sons.
- Epanechnikov, V.A. (1969). Nonparametric estimation of a multivariate probability density. *Theory Probab. Appl.*, 14, 153–158.
- Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 21, 826–838.
- Finklestein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42, 845–854.
- Fisher, R.A. (1922a). On the interpretation of χ^2 from contingency tables and the calculation of P. *J. Roy. Statist. Soc.*, 85, 87–94.
- Fisher, R.A. (1922b). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London A*, 222, 309–368.
- Fisher, R.A. (1925). *Statistical Methods for Research Workers*. (14th ed., 1970). Edinburgh: Oliver & Boyd.
- Fisher, R.A. (1935). *The Design of Experiments*. (8th ed., 1966). Edinburgh: Oliver & Boyd.
- Fisher, R.A. (1956). *Statistical Methods for Scientific Inference*. Edinburgh: Oliver & Boyd.
- Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76, 378–382.
- Fleiss, J.L. (1979). Confidence intervals for the odds ratio in case-control studies: the state of the art. *J. Chronic Dis.*, 32, 69–77.

- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons.
- Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons.
- Fleiss, J.L. and Cohen, J. (1973). The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. *Ed. Psychol. Meas.*, 33, 613–619.
- Fleiss, J.L., Cohen, J. and Everitt, B.S. (1969). Large-sample standard errors of Kappa and weighted Kappa. *Psychol. Bull.*, 72, 323–327.
- Fleiss, J.L., Nee, J.C.M. and Landis, J.R. (1979). Large sample variance of Kappa in the case of different sets of raters. *Psychol. Bull.*, 86, 974–977.
- Fleiss, J.L., Bigger, J.T., McDermott, M., Miller, J.P., Moon, T., Moss, A.J., Oakes, D., Rolnitzky, L.M. and Therneau, T.M. (1990). Nonfatal myocardial infarction is, by itself, an inappropriate end point in clinical trials in cardiology. *Circulation*, 81, 684–685.
- Fleiss, J.L., Levin, B., and Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*, 3rd ed., Hoboken, NJ: John Wiley & Sons.
- Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: John Wiley & Sons.
- Freedman, D.A. (1983). A note on screening regression equations. *Amer. Statist.*, 37, 152–155.
- Freedman, L.S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statist. Med.*, 1, 121–129.
- Freedman, L.S. and Pee, D. (1989). Return to a note on screening regression equations. *Amer. Statist.*, 43, 279–282.
- Freireich, E.J., Gehan, E.A., Frei III, E., Schroeder, L.R., Wolman, L.J., Anbari, R., Burgert, E.O., Mills, S.D., Pinkel, D., Selawry, O.S., Moon, J.H., Gendel, B.R., Spurr, C.L., Storrs, R., Haurani, F., Hoogstraten, B. and Lee, S. (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy. *Blood*, 21, 699–716.
- Frick, H. (1995). Comparing trials with multiple outcomes: the multivariate one-sided hypothesis with unknown covariances. *Biomet. J.*, 37, 909–917.
- Frome, E.L. (1983). The analysis of rates using Poisson regression models. *Biometrics*, 39, 665–674.
- Frome, E.L. and Checkoway, H. (1985). Epidemiologic programs for computers and calculators: use of Poisson regression models in estimating incidence rates and ratios. *Amer. J. Epidemiol.*, 121, 309–323.
- Gail, M.H. (1973). The determination of sample sizes for trials involving several independent 2×2 tables. *J. Chronic Dis.*, 26, 669–673.
- Gail, M.H. (1978). The analysis of heterogeneity for indirect standardized mortality ratios. *J. Roy. Statist. Soc. A*, 141, 224–234.

- Gail, M.H., Santner, T.J. and Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, 36, 255–266.
- Gail, M.H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrics*, 71, 431–444.
- Gart, J.J. (1971). The comparison of proportions: a review of significance tests, confidence intervals and adjustments for stratification. *Rev. Int. Statist. Inst.*, 39, 16–37.
- Gart, J.J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2×2 tables. *Biometrika*, 72, 673–677.
- Gart, J.J. and Tarone, R.E. (1983). The relation between score tests and approximate UMPU tests in exponential models common in biometry. *Biometrics*, 39, 781–786.
- Gart, J.J. and Zweifel, J.R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, 54, 181–187.
- Gastwirth, J.L. (1966). On robust procedures. *J. Amer. Statist. Assoc.*, 61, 929–948.
- Gastwirth, J.L. (1985). The use of maximum efficiency robust tests in combining contingency tables and survival analysis. *J. Amer. Statist. Assoc.*, 80, 380–384.
- Gastwirth, J.L. and Greenhouse, S.W. (1995). Biostatistical concepts and methods in the legal setting. *Statist. Med.*, 14, 1641–1653.
- Gaynor, J.J., Fener, E.J., Tan, C.C., Wu, D.H., Little, C.R., Straus, D.J., Clarkson, B.D. and Brennan, M.F. (1993). On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *J. Amer. Statist. Assoc.*, 88, 400–409.
- Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, 52, 203–223.
- George, S.L. and Desu, M.M. (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *J. Chronic Dis.*, 27, 15–29.
- Gill, R.D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tracts, 124. Amsterdam: Mathematisch Centrum.
- Gill, R.D. (1984). Understanding Cox's regression model: a martingale approach. *J. Amer. Statist. Assoc.*, 79, 441–447.
- Goodman, L.A. and Kruskal, W.H. (1972). Measures of association for cross classifications: IV. Simplification of asymptotic variances. *J. Amer. Statist. Assoc.*, 67, 415–421.
- Gray, R.J. (1988). A class of K -sample tests for comparing the cumulative incidence of a competing risk. *Ann. Statist.*, 16, 1141–1154.
- Greenland, S. (1984). A counterexample to the test-based principle of setting confidence limits. *Am. J. Epidemiol.*, 120, 4–7.
- Greenwood, M.A. (1926). Report on the natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33, 1–26. London: H. M. Stationery Office.

- Grizzle, J.E. (1967). Continuity correction in the χ^2 test for 2×2 tables. *Amer. Statist.*, 21, 28–32.
- Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25, 489–503.
- Guenther, W.C. (1977). Power and sample size for approximate chi-square tests. *Amer. Statist.*, 31, 83–85.
- Guilbaud, O. (1983). On the large-sample distribution of the Mantel-Haenszel odds-ratio estimator. *Biometrics*, 39, 523–525.
- Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. New York: Academic Press.
- Haldane, J.B.S. (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Human Genet.*, 20, 309–311.
- Halperin, M. (1977). Re: Estimability and estimation in case-referent studies. (Letter). *Am. J. Epidemiol.*, 105, 496–498.
- Halperin, M., Ware, J.H., Byar, D.P., Mantel, N., Brown, C.C., Koziol, J., Gail, M.H. and Green, S.B. (1977). Testing for interaction in an $I \times J \times K$ contingency table. *Biometrika*, 64, 271–275.
- Harrell, F.E. (1986). The PHGLM procedure. In *SAS Supplemental Library User's Guide*, Version 5. Cary, NC: SAS Institute, Inc.
- Harrington, D.P. and Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69, 553–566.
- Harris, M.I., Hadden, W.C., Knowler, W.C. and Bennett, P.H. (1987). Prevalence of diabetes and impaired glucose tolerance and plasma glucose levels in US population aged 20–74 yr. *Diabetes*, 36, 523–534.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, 72, 320–338.
- Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. New York: Chapman&Hall.
- Hauck, W.W. (1979). The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics*, 35, 817–819.
- Hauck, W.W. (1989). Odds ratio inference from stratified samples. *Comm. Statist. A*, 18, 767–800.
- Hauck, W.W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *J. Amer. Statist. Assoc.*, 72, 851–853.
- Helland, I.S. (1987). On the interpretation and use of R^2 in regression analysis. *Biometrics*, 43, 61–69.
- Higgins, J.E. and Koch, G.G. (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *Int. Statist. Rev.*, 45, 51–62.
- Hillis, S.L. and Davis, C.S. (1994). A simple justification of the iterative fitting procedure for generalized linear models. *Am. Statistician*, 48, 288–289.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800–802.

- Holford, T.R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics*, 36, 299–306.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6, 65–70.
- Hommel G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75, 383–386.
- Hosmer, D.W. and Lemeshow, S. (2004). *Applied Logistic Regression*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Hsieh, F.Y. and Lavori, P.W. (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Control. Clin. Trials*, 21, 552–560.
- Hsieh, F.Y., Bloch, D.A. and Larsen, M.D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statist. Med.*, 17, 1623–1634.
- Hu, M.X. and Lachin, J.M. (2003). Corrections for bias in maximum likelihood parameter estimates due to nuisance parameters. *Comm. Statist.*, 32, 619–639.
- Huber, P.J. (1967). The behavior of maximum likelihood estimators under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Neyman, J. (ed.), 221–233. Berkeley, CA: University of California Press.
- Irwin, J.O. (1935). Tests of significance for differences between percentages based on small numbers. *Metron*, 12, 83–94.
- Irwin, J.O. (1949). A note on the subdivision of χ^2 into components. *Biometrika*, 36, 130–134.
- Johansen, S. (1983). An extension of Cox's regression model. *Int. Statist. Rev.*, 51, 165–174.
- Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60, 267–278.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.
- Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53, 457–481.
- Karlin, S. and Taylor, H.M. (1975). *A First Course in Stochastic Processes*, 2nd ed. New York: Academic Press.
- Karon, J.M. and Kupper, L.L. (1982). In defense of matching. *Am. J. Epidemiol.*, 116, 852–866.
- Katz, D., Baptista, J., Azen, S.P. and Pike, M.C. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, 34, 469–474.
- Kay, R. (1982). The analysis of transition times in multistate stochastic processes using proportional hazard regression models. *Comm. Statist. A*, 11, 1743–1756.
- Kelsey, J.L., Whittemore, A.S., Evans, A.S. and Thompson, W.D. (1996). *Methods in Observational Epidemiology*, 2nd ed. New York: Oxford University Press.

- Kendall, Sir M. and Stuart, A. (1979). *The Advanced Theory of Statistics*, Vol. 2, 4th ed. New York: Macmillan.
- Kenny, S.D., Aubert, R.E. and Geiss, L.S. (1995). Prevalence and incidence of non-insulin-dependent diabetes. In *Diabetes in America*, 2nd ed., National Diabetes Data Group, 47–67. NIH Publication No. 95-1468. Bethesda, MD: The National Institutes of Health.
- Kent, J.T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69, 19–27.
- Kent, J.T. and O'Quigley, J. (1988). Measure of dependence for censored survival data. *Biometrika*, 75, 525–534.
- Kimball, A.W. (1954). Short-cut formulas for the exact partition of χ^2 in contingency Tables. *Biometrics*, 10, 452–458.
- Kleinbaum, D.G., Kupper, L.L. and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. New York: Van Nostrand Reinhold.
- Koch, G.G., McCanless, I. and Ward, J.F. (1984). Interpretation of statistical methodology associated with maintenance trials. *Am. J. Med.*, 77(Suppl. 5B), 43–50.
- Korn, E.L. (1984). Estimating the utility of matching in case-control studies. *J. Chronic Dis.*, 37, 765–772.
- Korn, E.L. and Simon R.M. (1990). Measures of explained variation for survival data. *Statist. Med.*, 9, 487–503
- Korn, E.L. and Simon, R.M. (1991). Explained residual variation, explained risk, and goodness of fit. *Amer. Statist.*, 45, 201–206.
- Kraemer, H.C., Vyjeyanthi, S.P. and Noda, A. (2004). Kappa coefficients in medical research. In *Tutorials in Biostatistics, Vol. 1; Statistical Methods in Clinical Studies*. D'Agostino, R.B. (ed.). New York: John Wiley & Sons.
- Kruskal W.H. and Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.*, 47, 583–612.
- Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, 50, 403–418.
- Kupper, L.L. and Hafner, K.B. (1989). How appropriate are popular sample size formulas? *Amer. Statist.*, 43, 101–105.
- Kupper, L.L., Karon, J.M., Kleinbaum, D.G., Morgenstern, H. and Lewis, D.K. (1981). Matching in epidemiologic studies: validity and efficiency considerations. *Biometrics*, 37, 271–292.
- Lachin, J.M. (1977). Sample size determinations for $r \times c$ comparative trials. *Biometrics*, 33, 315–324.
- Lachin, J.M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Control. Clin. Trials*, 2, 93–113.
- Lachin, J.M. (1992a). Some large sample distribution-free estimators and tests for multivariate partially incomplete observations from two populations. *Statist. Med.*, 11, 1151–1170.

- Lachin, J.M. (1992b). Power and sample size evaluation for the McNemar test with application to matched case-control studies. *Statistics in Medicine*, 11, 1239–1251.
- Lachin, J.M. (1996). Distribution-free marginal analysis of repeated measures. *Drug Inform. J.*, 30, 1017–1028.
- Lachin, J.M. (1998). Sample size determination. In *Encyclopedia of Biostatistics*, Armitage, P. and Colton, T. (eds.), 3892–3903. New York: John Wiley & Sons.
- Lachin, J.M. (2004). The role of measurement reliability in clinical trials. *Clin. Trials*, 1, 553–566.
- Lachin, J.M. (2005). Maximum information designs. *Clinical Trials*, 2, 453–464.
- Lachin, J.M. (2008). Sample size evaluation for a multiply matched casecontrol study using the score test from a conditional logistic (discrete Cox PH) regression model. *Statist. Med.* 27, 25092523.
- Lachin, J.M. and Bautista, O.M. (1995). Stratified-adjusted versus unstratified assessment of sample size and power for analyses of proportions. In *Recent Advances in Clinical Trial Design and Analysis*, Thall, P.F. (ed.), 203–223. Boston: Kluwer Academic.
- Lachin, J.M. and Foulkes, M.A. (1986). Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, non-compliance and stratification. *Biometrics*, 42, 507–519.
- Lachin, J.M. and Wei, L.J. (1988). Estimators and tests in the analysis of multiple nonindependent 2×2 tables with partially missing observations. *Biometrics*, 44, 513–528.
- Lachin, J.M., Lan, S.L. and the Lupus Nephritis Collaborative Study Group (1992). Statistical considerations in the termination of a clinical trial with no treatment group difference: the Lupus Nephritis Collaborative Study. *Control. Clin. Trials*, 13, 62–79.
- Lagakos, S.W. (1978). A covariate model for partially censored data subject to competing causes of failure. *Appl. Statist.*, 27, 235–241.
- Lagakos, S.W. (1988). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika*, 75, 156–160.
- Lagakos, S.W. and Schoenfeld, D. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics*, 40, 1037–1048.
- Lagakos, S.W., Limm L.L.-Y. and Robins, J.N. (1990). Adjusting for early treatment termination in comparative clinical trials. *Statist. Med.*, 9, 1417–1424.
- Laird, N. and Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.*, 76, 231–240.
- Lakatos E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, 44, 229–241.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1–14.
- Lan, K.K.G. and Lachin, J.M. (1995). Martingales without tears. *Lifetime Data Anal.*, 1, 361–375.

- Lan, K.K.G. and Wittes, J.T. (1985). Rank tests for survival analysis: a comparison by analogy with games. *Biometrics*, 41, 1063–1069.
- Lancaster, H.O. (1949). The derivation and partition of χ^2 in certain discrete distributions. *Biometrika*, 36, 117–129.
- Lancaster, H.O. (1950). The exact partition of χ^2 and its application to the problem of pooling of small expectations. *Biometrika*, 37, 267–270.
- Landis, J.R., Heyman, E.R. and Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *Int Statist. Rev.*, 46, 237–254.
- Landis, J.R. and Koch, G.G. (1997). A one-way components of variance model for categorical data. *Biometrics*, 33, 671–679.
- Laupacis, A., Sackett, D.L. and Roberts, R.S. (1988). An assessment of clinically useful measures of the consequences of treatment. *N. Engl. J. Med.*, 318, 1728–1733.
- Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons.
- Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *Can. J. Statist.*, 15, 209–225.
- Lee, E.T. (1992). *Statistical Methods for Survival Data Analysis*, 2nd ed. New York: John Wiley & Sons.
- Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis*, Klein, J.P. and Goel, P.K. (eds.). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. London: Chapman & Hall.
- Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd ed. London: Chapman & Hall.
- Lehmann, E.L. (1998). *Elements of Large-Sample Theory*. New York: Springer-Verlag.
- Leung, H.K. and Kupper, L.L. (1981). Comparison of confidence intervals for attributable risk. *Biometrics*, 37, 293–302.
- Levin, M.L. (1953). The occurrence of lung cancer in man. *Acta Unio Int. Contra Cancrum*, 19, 531–541.
- Lewis, E.J., Hunsicker, L.G., Lan, S.L., Rohde, R.D., Lachin, J.M. and the Lupus Nephritis Collaborative Study Group (1992). A controlled trial of plasmapheresis therapy in severe lupus nephritis. *N. Engl. J. Med.*, 326, 1373–1379.
- Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- Liang, K.Y. and Zeger, S.L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statist. Sci.*, 10, 158–173.

- Lin, D.Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J. Amer. Statist. Assoc.*, 86, 725–728.
- Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statist. Med.*, 13, 2233–2247.
- Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.*, 84, 1074–1078.
- Lin, D.Y. and Wei, L.J. (1991). Goodness-of-fit tests for the general Cox regression model. *Statist. Sinica*, 1, 1–17.
- Lin, D.Y., Wei, L.J. and Zing, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557–72.
- Lin, D.Y., Wei L.J., Yang I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. Roy. Statist. Soc. B*, 62, 711–730.
- Lin, J.S., and Wei, L.J. (1992). Linear regression analysis for multivariate failure time observations. *J. Amer. Statist. Assoc.*, 87, 1091–1097
- Lipsitz, S.H., Fitzmaurice, G.M., Orav, E.J. and Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 50, 270–278.
- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Liu, W. (1996). Multiple tests of a non-hierarchical finite family of hypotheses. *J. Roy. Statist. Soc. B*, 58, 455–461.
- Louis, T.A. (1981). Confidence intervals for a binomial parameter after observing no successes. *Amer. Statist.*, 35, 154.
- Lu, B., Preisser, J.S., Qaqish, B.F., Suchindran, C., Bangdiwala, S.I., and Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63, 935–941.
- Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*, 51, 524–532.
- Machin, D. and Campbell, M. J. (1987). *Statistical Tables for the Design of Clinical Trials*. Oxford, UK: Blackwell Scientific.
- Mack, T.M., Pike, M.C., Henderson, B.E., Pfeffer, R.I., Gerkins, V.R., Arthus, B.S. and Brown, S.E. (1976). Estrogens and endometrial cancer in a retirement community. *N. Engl. J. Med.*, 294, 1262–1267.
- MacMahon, B. and Pugh, T.F. (1970). *Epidemiology: Principles and Methods*. Boston: Little, Brown.
- Madalla, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.
- Magee, L. (1990). R^2 measures based on Wald and likelihood ratio joint significance tests. *Amer. Statist.*, 44, 250–253.
- Makuch, R.W. and Simon, R.M. (1978). Sample size requirements for evaluating a conservative therapy. *Cancer Treat Rep.*, 62, 1037–40.
- Mallows, C.L. (1973). Some comments on C_p . *Technometrics*, 15, 661–675.

- Mann, H.B. and Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18, 50-60.
- Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J. Amer. Statist. Assoc.*, 58, 690-700.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, 50, 163-170.
- Mantel, N. (1967). Ranking procedures for arbitrarily restricted observations. *Biometrics*, 23, 65-78.
- Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, 12, 591-612.
- Mantel, N. (1974). Comment and a suggestion. *J. Amer. Statist. Assoc.*, 69, 378-380.
- Mantel, N. (1987). Understanding Wald's test for exponential families. *Amer. Statist.*, 41, 147-148.
- Mantel, N. and Greenhouse, S.W. (1968). What is the continuity correction? *Amer. Statist.*, 22, 27-30.
- Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, 22, 719-748.
- Mantel, N., Brown, C.C. and Byar, D.P. (1977). Tests for homogeneity of effect in an epidemiologic investigation. *Am. J. Epidemiol.*, 106, 125-129.
- Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special references to ordered analysis of variance. *Biometrika*, 63, 655-600.
- Martin, D.C. and Katti, S.K. (1965). Fitting of certain contagious distributions to some available data by the maximum likelihood method. *Biometrics*, 21, 34-48.
- Marubini, E. and Valsecchi, M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. New York: John Wiley & Sons.
- McCullagh, P. (1980). Regression models for ordinal data, *J. Roy. Statist. Soc. B*, 42, 109-142.
- McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.*, 11, 59-67.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman&Hall.
- McHugh, R.B. and Le, C.T. (1984). Confidence estimation and the size of a clinical trial. *Control. Clin. Trials*, 5, 157-163.
- McKinlay, S.M. (1978). The effect of non-zero second order interaction on combined estimators of the odds-ratio. *Biometrika*, 65, 191-202.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157.
- Mehrotra, K.G., Michalek, J.E. and Mihalko, D. (1982). A relationship between two forms of linear rank procedures for censored data. *Biometrika*, 69, 674-676.
- Mehta, C. and Patel, N. (1999). *StatXact 4 for Windows*. Cambridge, MA: Cytel Software Corporation.
- Meier, P. (1953). Variance of a weighted mean. *Biometrics*, 9, 59-73.

- Meng, R.C. and Chapman, D.G. (1966). The power of the chi-square tests for contingency tables. *J. Amer. Statist. Assoc.*, 61, 965–975.
- Miettinen, O.S. (1968). The matched pairs design in the case of all-or-none responses. *Biometrics*, 24, 339–352.
- Miettinen, O.S. (1969). Individual matching with multiple controls in the case of all-or-none responses. *Biometrics*, 25, 339–55.
- Miettinen, O.S. (1970). Matching and design efficiency in retrospective studies. *Am. J. Epidemiol.*, 91, 111–118.
- Miettinen, O.S. (1974a). Comment. *J. Amer. Statist. Assoc.*, 69, 380–382.
- Miettinen, O.S. (1974b). Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am. J. Epidemiol.*, 99, 325.
- Miettinen, O.S. (1976). Estimability and estimation in case-referent studies. *Am. J. Epidemiol.*, 103, 226–235.
- Miller, A.J. (1984). Selection of subsets of regression variables. *J. Roy. Statist. Soc. A*, 147, 389–425.
- Mogensen, C.E. (1984). Microalbuminuria predicts clinical proteinuria and early mortality in maturity-onset diabetes. *N. Engl. J. Med.*, 310, 356–360.
- Moore, D.F. (1986). Asymptotic properties of moment estimators for over-dispersed counts and proportions. *Biometrika*, 73, 583–588.
- Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc.*, 78, 47–55.
- Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.
- Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc., 135*, 370–384.
- Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14, 945–965.
- Newcombe, R.G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statist. Med.*, 17, 857–872.
- Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics*, Grenander, U. (ed.), 213–234. Stockholm, Sweden: Almqvist & Wiksell.
- Noether, G.E. (1955). On a theorem of Pitman. *Ann. Math. Statist.*, 26, 64–68.
- Noether, G.E. (1987). Sample size determination for some common nonparametric tests. *J. Amer. Statist. Assoc.*, 82, 645–647.
- O'Brien, R.G. and Castelloe, J.M. (2006). Exploiting the link between the Wilcoxon–Mann–Whitney test and a simple odds statistic. In *Proceedings of the Thirty-first Annual SAS Users Group International Conference*, Paper 209–31. Cary, NC: SAS Institute Inc.
- Odeh, R.E. and Fox, M. (1991). *Sample Size Choice: Charts for Experiments with Linear Models*, 2nd ed. New York: Marcel Dekker.

- Odell, P.M., Anderson, K.M. and D'Agostino, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48, 951–959.
- O'Quigley, J. and Flandre, P. (1994). Predictive capability of proportional hazards regression. *Proc. Natl. Acad. Sci. U.S.A.*, 91, 2310–2314.
- O'Quigley, J., Flandre, P. and Reiner, E. (1999). Large sample theory for Schemper's measures of explained variation in the Cox regression model. *Statistician*, 48, 53–62.
- Palesch, Y.Y. and Lachin, J.M. (1994). Asymptotically distribution-free multivariate rank tests for multiple samples with partially incomplete observations. *Statist. Sinica*, 4, 373–387.
- Palta, M. and Amini, S.B. (1985). Consideration of covariates and stratification in sample size determination for survival time studies. *J. Chronic Dis.*, 38, 801–809.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London Edinburg Dublin Philos. Mag. J. Sci.*, 50, 157–175.
- Pepe, M.S. (1991). Inference for events with dependent risks in multiple endpoint studies. *J. Amer. Statist. Assoc.*, 86, 770–778.
- Pepe, M.S. and Mori, M. (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statist. Med.*, 12, 737–751.
- Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *Ann. Math. Statist.*, 40, 549–567.
- Peterson, A.V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.*, 72, 854–858.
- Peterson, B. and Harrell, F.E. (1990). Partial proportional odds models for ordinal response variables. *Appl. Statist.*, 39, 205–217.
- Peto, J. (1984). The calculation and interpretation of survival curves. In *Cancer Clinical Trials: Methods and Practice*, Buyse, M.E., Staquet, M.J. and Sylvester, R.J. (eds.). Oxford, UK: Oxford University Press.
- Peto, R. (1972). Contribution to the discussion of paper by D. R. Cox. *J. Roy. Statist. Soc. B*, 34, 205–207.
- Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.*, 22, 86–91.
- Peto, R. (1987). Why do we need systematic overviews of randomized trials? *Statist. Med.*, 6, 233–240.
- Peto, R. and Lee, P. (1983). Weibull distributions for continuous carcinogenesis experiments. *Biometrics*, 29, 457–470.
- Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J. Roy. Statist. Soc., A*, 135, 185–206.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, V., Mantel, N., McPherson, K., Peto, J. and Smith, P.G. (1976). Design and analysis

- of randomised clinical trials requiring prolonged observation of each patient: introduction and design. *Br. J. Cancer*, 34, 585–612.
- Pettigrew, H.M., Gart, J.J. and Thomas, D.G. (1986). The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika*, 73, 425–435.
- Pike, M.C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, 22, 142–61.
- Pirart, J. (1978a). Diabetes mellitus and its degenerative complications: a prospective study of 4,400 patients observed between 1947 and 1973. *Diabetes Care*, 1, 168–188.
- Pirart, J. (1978b). Diabetes mellitus and its degenerative complications: a prospective study of 4,400 patients observed between 1947 and 1973. *Diabetes Care*, 1, 252–263.
- Pitman, E.J.G. (1948). *Lecture Notes on Nonparametric Statistics*. New York: Columbia University.
- Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.*, 9, 705–724.
- Prentice, R.L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika*, 60, 279–288.
- Prentice, R.L. (1978). Linear rank tests with right-censored data. *Biometrika*, 65, 167–179.
- Prentice, R.L. and Gloeckler, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34, 57–67.
- Prentice, R.L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics*, 35, 861–867.
- Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Flournoy, N., Farewell, V.T. and Breslow, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34, 541–554.
- Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the regression analysis of multivariate failure time. *Biometrika*, 68, 373–379.
- Radhakrishna, S. (1965). Combination of results from several 2×2 contingency tables. *Biometrics*, 21, 86–98.
- Ramlau-Hansen, H. (1983a). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, 11, 453–466.
- Ramlau-Hansen, H. (1983b). The choice of a kernel function in the graduation of counting process intensities. *Scand. Actuar. J.*, 165–182.
- Rao, C.R. (1963). Criteria of estimation in large samples. *Sankhya A*, 25, 189–206.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Application*, 2nd ed. New York: John Wiley & Sons.
- Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, 35, 1–20.
- Robins, J.N., Breslow, N.E., and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, 42, 311–323.

- Robins, J.N., Greenland, S. and Breslow, N.E. (1986). A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am. J. Epidemiol.*, 124, 719–723.
- Rochon, J.N. (1989). The application of the GSK method to the determination of minimum sample sizes. *Biometrics*, 45, 193–205.
- Ross, S.M. (1983). *Stochastic Processes*. New York: John Wiley & Sons.
- Rothman, K.J. (1986). *Modern Epidemiology*. Boston: Little, Brown.
- Royall, R.M. (1986). Model robust inference using maximum likelihood estimators. *Int. Statist. Rev.*, 54, 221–226.
- Rubenstein L.V., Gail, M.H. and Santner, T.J. (1981). Planning the duration of a comparative clinical trial with losses to follow-up and a period of continued observation. *J. Chronic Dis.*, 34, 469–479.
- Sahai, H. and Khurshid, A. (1995). *Statistics in Epidemiology*. Boca Raton, FL: CRC Press.
- SAS Institute (1995). *Logistic Regression Examples Using the SAS System, Version 6*. Cary, NC: SAS Institute, Inc.
- SAS Institute (2008a). *Base SAS 9.2 Procedures Guide: Statistical Procedures*. Cary, NC: SAS Institute, Inc.
- SAS Institute (2008b). *SAS/STAT 9.2 User's Guide*. Cary, NC: SAS Institute, Inc.
- Schemper M. (1990). The explained variation in proportional hazards regression. *Biometrika*, 77, 216–218. (Correction: 81, 631).
- Schemper, M. (1992). Further results on the explained variation in proportional hazards regression. *Biometrika*, 79, 202–204.
- Schemper, M. and Stare, J. (1996). Explained variation in survival analysis. *Statist. Med.*, 15, 1999–2012.
- Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. New York: Oxford University Press.
- Schoenfeld D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68, 316–319.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 239–241.
- Schoenfeld, D. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39, 499–503.
- Schrek, R., Baker, L.A., Ballard, G.P. and Dolgoff, S. (1950). Tobacco smoking as an etiologic factor in disease. I. Cancer. *Cancer Res.*, 10, 49–58.
- Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.*, 15, 657–680.
- Schuster, J.J. (1990). *CRC Handbook of Sample Size Guidelines for Clinical Trials*. Boca Raton, FL: CRC Press.
- Scrucca, L., Santucci, A. and Aversa, F. (2007). Competing risk analysis using R: an easy guide for clinicians. *Bone Marrow Transplant.*, 40, 381–387.

- Selvin, S. (1996). *Statistical Analysis of Epidemiologic Data*, 2nd edition. New York: Oxford University Press.
- Serfling, R.J. (1980). *Approximation Theorems of Statistics*. New York: John Wiley & Sons.
- Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27, 379–423 and 623–656.
- Shoukri, M.M. (2004). *Measures of Interobserver Agreement*. New York: Chapman&Hall/CRC.
- Slager, S.L. and Schaid, D.J. (2001). Case-control studies of genetic markers: power and sample size approximation for Armitage's test of trend. *Human Hered.*, 52, 149–153.
- Snedecor, G.W. and Cochran, W.G. (1967). *Statistical Methods*, 6th ed. Ames, IA: The Iowa State University Press.
- Sparling, Y.H., Younes, N., Lachin, J.M. and Bautista, O.M. (2006). Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics*, 7, 599–614.
- Spearman C. (1904). The proof and measurement of association between two things. *Amer. J. Psychol.*, 15, 72–101.
- Starmer, C.F., Grizzle, J.E. and Sen, P.K. (1974). Comment. *J. Amer. Statist. Assoc.*, 69, 376–378.
- Steffes, M.W., Chavers, B.M., Bilous, R.W. and Mauer, S.M. (1989). The predictive value of microalbuminuria. *Amer. J. Kidney Dis.*, 13, 25–28.
- Stokes, M.E., Davis, C.S. and Koch, G.G. (2000). *Categorical Data Analysis Using the SAS System*, 2nd ed. Cary, NC: SAS Institute, Inc.
- Stuart, A. (1955). A test for homogeneity of the marginal distribution in a two-way classification. *Biometrika*, 42, 412–416.
- Sun, J., Zhao, Q. and Zhao X. (2005). Generalized log-rank tests for interval-censored failure time data. *Scand. J. Statist.*, 32, 49–57.
- Sun, X. and Yang, Z. (2008). Generalized McNemar's test for homogeneity of the marginal distributions. *SAS Global Forum 2008: Statistics and Data Analysis*, Paper 382–208.
- Tang, D.I., Gnecco, C. and Geller, N.L. (1989). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika*, 76, 577–583.
- Tang, Y. (2009). Comments on "Sample size evaluation for a multiply matched casecontrol study using the score test from a conditional logistic (discrete Cox PH) regression model". *Statist. Med.*, 28, 173179.
- Tarone, R.E. (1985). On heterogeneity tests based on efficient scores. *Biometrika*, 72, 91–95.
- Tarone, R.E. and Ware J.H. (1977). On distribution free tests for equality of survival distributions. *Biometrika*, 64, 156–160.
- Thall, P.F. and Lachin, J.M. (1986). Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. *Stat. Med.*, 5, 73–83.

- Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with over-dispersion. *Biometrics*, 46, 657-671.
- Thall, P.F., Russell, K.E. and Simon, R.M. (1997). Variable selection in regression via repeated data splitting. *J. Comput. Graph. Statist.*, 6, 416-434.
- Thall, P.F., Simon, R.M. and Grier, D.A. (1992). Test-based variable selection via cross-validation. *J. Comput. Graph. Statist.*, 1, 41-61.
- Therneau, T.M., Grambsch, P.M. and Fleming, T.R. (1990). Martingale hazards regression models and the analysis of censored survival data. *Biometrika*, 77, 147-160.
- Thisted, R.A. (1988). *Elements of Statistical Computing*. New York: Chapman&Hall.
- Thomas, D.G. and Gart, J.J. (1977). A table of exact confidence limits for differences and ratios of two proportions and their odds ratios. *J. Amer. Statist. Assoc.*, 72, 73-76.
- Thomas, D.C. and Greenland, S. (1983). The relative efficiencies of matched and independent sample designs for case-control studies. *J. Chronic Dis.*, 36, 685-697.
- Tocher, K.D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika*, 37, 130-144.
- Truett, J., Cornfield, J. and Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chronic Dis.*, 20, 511-524.
- Tsiatis, A.A. (1981). A large sample study of Cox's regression model. *Ann. Statist.*, 9, 93-108.
- Turnbull, B.W. (1976). The empirical distribution function with arbitrarily censored and truncated data. *J. Roy. Statist. Soc. B*, 38, 290-295.
- University Group Diabetes Program (UGDP) (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. *Diabetes*, 19(Suppl. 2), Appendix A, 816-830.
- U.S. Surgeon General (1964). *Smoking and Health*. Publication (PHS) 1103. U.S. Department of Health, Education, and Welfare.
- U.S. Surgeon General. (1982). *The Health Consequences of Smoking: Cancer*. Publication (PHS) 82-50179. Rockville, MD: U.S. Department of Health and Human Services.
- Vaeth, M. (1985). On the use of Wald's test in exponential families. *Int. Statist. Rev.*, 53, 199-214.
- van Elteren, P.H. (1960). On the combination of independent two-sample tests of Wilcoxon. *Bull. Int. Statist. Inst.*, 37, 351-361.
- Wacholder, S. and Weinberg, C.R. (1982). Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: power considerations. *Biometrics*, 38, 801-812.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.*, 54, 426-482.

- Wallenstein, S. and Wittes, J.T. (1993). The power of the Mantel-Haenszel test for grouped failure time data. *Biometrics*, 49, 1077–1087.
- Walter, S.D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika*, 62, 371–375.
- Walter, S.D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics*, 32, 829–849.
- Walter, S.D. (1978). Calculation of attributable risks from epidemiological data. *Int. J. Epidemiol.*, 7, 175–182.
- Walter, S.D. (1980). Matched case-control studies with a variable number of controls per case. *Appl. Statist.*, 29, 172–9.
- Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 63, 27–32.
- Wei, L.J. and Glidden, D.V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Stat. in Med.*, 16, 833–839.
- Wei, L.J. and Lachin, J.M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J. Amer. Statist. Assoc.*, 79, 653–661.
- Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.*, 84, 1065–1073.
- White, A.A., Landis, J.R., and Cooper, M.M. (1982). A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Int. Statist. Rev.*, 50, 27–34.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.
- Whitehead, A. (2002). *Meta-analysis of Controlled Clinical Trials*. Hoboken, NJ: John Wiley & Sons.
- Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statist. Med.*, 10, 1665–1677.
- Whitehead, J. (1989). The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments. *Statist. Med.*, 8, 1439–1454.
- Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*, 2nd ed. New York: Ellis Horwood.
- Whittemore, A.S. (1981). Sample size for logistic regression with small response probability. *J. Amer. Statist. Assoc.*, 76, 27–32.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bull.*, 1, 80–83.
- Wilks, S.S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons.
- Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.*, 22, 209–212.

- Wittes, J.T. and Wallenstein, S. (1987). The power of the Mantel-Haenszel test. *J. Amer. Statist. Assoc.*, 82, 400, 1104–1109.
- Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann. Human Genet.*, 19, 251–253.
- Wright, S.P. (1992). Adjusted *p*-values for simultaneous inference. *Biometrics*, 48, 1005–1013.
- Younes, N. and Lachin, J.M. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, 53, 1199–1211.
- Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, T. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog. Cardiovasc. Dis.*, 27, 335–371.
- Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121–130.
- Zelen, M. (1971). The analysis of several 2×2 contingency tables. *Biometrika*, 58, 129–137.
- Zhao, Y.D., Rahardja, D. and Qu, Y. (2008). Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties. *Statist. Med.*, 27, 462–468.
- Zuckerman, D.M., Kasl, S.V. and Ostfeld, A.M. (1984). Psychosocial predictors of mortality among the elderly poor: the role of religion, well-being, and social contacts. *Am. J. Epidemiol.*, 119, 419–423.

Author Index

- Aalen, O., 434–435, 449, 456, 502, 504, 593
Agresti, A., 42, 73, 235, 593
Akaike, H., 307, 593
Alioum, A., 495, 593
Allison, T., 237, 596
Altschuler, B., 436, 593
Amato, D.A., 499, 606
Amini, S.B., 489, 610
Anbari, R., 520, 523, 600
Andersen, P.K., 429, 456, 459, 463, 499,
 503–504, 506–508, 593
Anderson, J.A., 308, 593
Anderson, J.R., 453, 594
Anderson, K.M., 496–497, 610
Anderson, T.W., 152, 594
Andrews, D.F., 595
Anscombe, F.J., 31, 594
Aranda-Ordaz, F.J., 496, 594
Armitage, P., 74, 453–454, 594, 605, 610
Arthus, B.S., 222, 607
Aubert, R.E., 14, 604
Aversa, F., 494, 612
Azen, S.P., 24, 603
Bailar, J.C., 5, 594
Baker, L.A., 206, 612
Ballard, G.P., 206, 612
Bancroft, T.A., 164, 594
Bangdiwala, S.I., 364, 607
Baptista, J., 24, 603
Barnard, G.A., 33–35, 594
Bartlett, M.S., 343, 594
Bautista, O.M., 141, 188, 497, 605, 613
Beach, M.L., 141, 594
Bean, S.J., 503, 594
Bennett, P.H., 7, 14, 602
Bennett, S., 451, 454, 523, 594
Bernstein, L., 453, 594
Bhapkar, V.P., 235, 594
Bickel, P., 535, 544, 594
Bigger, J.T., 495, 600
Bilous, R.W., 8, 595, 613
Birch, M.W., 30, 131, 183, 257, 594
Bishop, Y.M.M., 235, 594
Blackwelder, W.C., 96–97, 594
Bloch D.A., 236
Bloch, D.A., 161, 321–322, 325, 594, 603
Blum, A.L., 122, 594
Boos, D., 591, 595
Borgan, O., 429, 463, 499, 503–504, 506, 593
Bowker, A.H., 234, 595
Breiman, L., 307, 595
Brennan, M.F., 492, 494, 527, 601
Breslow, N.E., 13, 125, 129, 135, 155–156, 218,
 222, 228–229, 243, 261, 283, 308, 337,
 341, 379–380, 402, 453–454, 467–468,
 471, 492, 595, 610–612
Bronowski, J., 1, 595
Bross, I.D.J., 64, 595
Brown, C.C., 154, 421, 601–602, 608
Brown, S.E., 222, 607

- Burgert, E.O., 520, 523, 600
 Buyse, M.E., 610
 Byar, D.P., 154, 200, 260–261, 291, 531, 595, 597, 602, 608
 Campbell, M.J., 85, 607
 Canner, P.L., 141, 595
 Castelhoe, J.M., 102, 107, 609
 Chapman, D.G., 100, 609
 Chavers, B.M., 8, 595, 613
 Checkoway, H., 422–423, 600
 Cicchetti, D.V., 237, 596
 Clarkson, B.D., 492, 494, 527, 601
 Clayton, D., 357, 463
 Clayton, D.G., 596
 Clopper, C.J., 16, 596
 Cochran, W.G., 41, 64, 74, 120, 125, 140, 153, 235, 247, 303, 388, 596, 613
 Cohen, J., 85, 235–238, 596, 600
 Collett, D., 376, 429, 435, 469, 596
 Collins, R., 182–183, 258, 596, 616
 Colton, T., 605
 Commenges, D., 495, 593
 Connell J.E., 596
 Connell, J.E., 223
 Connor, R.J., 223, 596
 Conover, W.J., 43–44, 596
 Cook, R.J., 56, 596
 Cooper, M.M., 235, 615
 Cornfield, J., 32, 40, 55, 73, 76, 204–206, 222, 287, 307, 493, 596–597, 614
 Coronary Drug Project Research Group, CDP, 307, 595
 Cox, D.R., 215, 250, 264–265, 280, 283, 382, 429, 450, 453–454, 456–457, 459, 466, 469, 507, 535, 558, 567, 597, 610
 Cramér, H., 69, 535, 543, 597
 Csaki, F., 593
 Cutler, S.J., 443, 597
 Cuzick, J., 463, 596
 D'Agostino, R.B., 496–497, 604, 610
 Davies, M., 239, 597
 Davis, C.S., 136, 145, 189, 283, 587, 602, 613
 Day, N.E., 13, 155–156, 222, 229, 260–261, 291, 308, 341, 379–380, 595, 597
 DCCT, 598
 Dean, C., 402, 597
 Deckert T., 7
 Deckert, T., 597
 Demidenko, E., 359–360, 597
 DerSimonian, R., 177, 359, 597
 Desu, M.M., 85, 485, 597, 601
 Diabetes Control and Complications Trial Research Group, 9–11, 65, 219, 231, 238, 245, 445, 482–483, 506, 509, 511, 598
 Diabetes Prevention Program Research Group, 71–72, 598
 Dick, T.D.S., 308, 598
 Diggle, P.J., 588, 598
 Dobson, A., 580, 598
 Doksum, K., 535, 544, 594
 Dolgoff, S., 206, 612
 Donner, A., 85, 313, 598, 602
 Dorn, H.F., 206, 599
 Dyke, G.V., 293, 599
 Early Breast Cancer Trialists' Collaborative Group, 599
 Ederer, F., 443, 597
 Efron, B., 248, 334, 343, 473, 573, 599
 Efroymson, M.A., 305, 599
 Ejigou, A., 215, 599
 Elandt-Johnson, R.C., 429, 444, 492, 599
 Ellis, E.N., 8, 595, 615
 Epanechnikov, V.A., 504, 599
 Evans, A.S., 13, 137, 603
 Everitt, B.S., 236, 238, 600
 Farewell, V.T., 492, 611
 Feigl, P., 496, 599
 Feinberg, S.E., 235
 Fener, E.J., 492, 494, 527, 601
 Fienberg, S.E., 594
 Finklestein, D.M., 495, 599
 Fisher, R.A., 33, 58–61, 247, 259, 265, 270–271, 599
 Fitzmaurice, G.M., 180, 607
 Flandre, P., 471, 610
 Fleiss, J.L., 13, 32, 177, 235–239, 495, 597, 599–600
 Fleming, T.R., 425, 427, 429, 449, 452, 470, 499, 504, 506, 532, 600, 602, 614
 Flournoy, N., 492, 611
 Foulkes, M.A., 484–486, 489, 605
 Fox, M., 85, 609
 Freedman, D.A., 305
 Freedman, L.S., 305, 484, 600
 Frei III, E., 520, 523, 600
 Freireich, E.J., 520, 523, 600
 Frick, H., 173, 600
 Frome, E.L., 394, 422–423, 600
 Gabriel, K.R., 70, 608
 Gail, M.H., 154, 183, 401, 421, 463, 485, 600–602, 612
 Gart, J.J., 30–32, 36, 80, 120, 131, 134, 263, 280, 601, 611, 614
 Gastwirth, J.L., 170–171, 373, 453, 601
 Gaynor, J.J., 601
 Gehan, E.A., 451, 453, 520, 523, 600–601
 Geiss, L.S., 14, 604
 Geller, N.L., 173, 613
 Gendel, B.R., 520, 523, 600
 George, S.L., 485, 601
 Gerkins, V.R., 222, 607

- Gill, R.D., 429, 434, 449, 456, 459, 463, 499, 503–504, 506–508, 593, 601
 Glidden, D.V., 512, 615
 Gloeckler, L.A., 465–466, 611
 Gnecco, C., 173, 613
 Goel, P.K., 606
 Goodman, L.A., 376, 601
 Grambsch, P.M., 470, 614
 Gray, R.J., 492, 494, 527, 601
 Green, S.B., 154, 602
 Greenhouse, S.W., 44, 373, 601, 608
 Greenland, S., 43, 128–129, 218, 227–228, 243, 261, 601, 611–612, 614
 Greenwood, M.A., 434, 601
 Grenander, U., 609
 Grier, D.A., 307, 614
 Grizzle, J.E., 43–44, 152, 235, 288, 602, 613
 Guenther, W.C., 100, 602
 Guilbaud, O., 129, 602
 Hadden, W.C., 7, 14, 602
 Haenszel, W., 39, 64, 120, 123, 126, 144, 217, 247, 608
 Hafner, K.B., 86, 604
 Hájek, J., 64, 112, 452, 602
 Haldane, J.B.S., 31, 602
 Halperin, M., 128, 154, 602
 Halvorsen, K.T., 341, 595
 Harrell, F.E., 358, 472, 602, 610
 Harrington, D.P., 425, 427, 429, 449, 452, 470, 499, 504, 506, 532, 600, 602
 Harris, M.I., 7, 14, 602
 Harville, D.A., 177, 602
 Hastie, T.J., 302, 304, 602
 Hauck, W.W., 128, 135, 313, 602
 Haurani, F., 520, 523, 600
 Helland, I.S., 471, 571, 602
 Henderson, B.E., 222, 607
 Herzberg, A.M., 595
 Heyman, E.R., 189, 606
 Higgins, J.E., 197, 602
 Hillis, S.L., 587, 602
 Hinkley, D.V., 248, 265, 466, 535, 558, 567, 597, 599
 Hochberg, Y., 70, 602
 Holford, T.R., 402, 603
 Holland, P.W., 235, 594
 Holm, S., 70, 603
 Hommel G., 70–71, 603
 Hoogstraten, B., 520, 523, 600
 Hosmer, D.W., 233, 283, 302, 304, 345–346, 603
 Howard, V., 453–454, 610
 Hsieh, F.Y., 321–322, 325, 350, 486, 488, 603
 Hu, M.X., 343, 603
 Huber, P.J., 574, 603
 Hunsicker, L.G., 490, 606
 Irwin, J.O., 33, 63, 603
 Johansen, S., 459, 603
 Johnson, N.L., 429, 444, 492, 599
 Juhl, E., 594
 Kalbfleisch, J.D., 429, 459, 467–469, 471, 473, 492, 603, 611
 Kannel, W., 307, 614
 Kaplan, E.L., 432–433, 603
 Karlin, S., 382, 603
 Karon, J.M., 226–227, 603–604
 Kasl, S.V., 142, 616
 Katti, S. K., 407
 Katti, S.K., 608
 Katz, D., 24, 603
 Kay, R., 507, 603
 Keiding, N., 429, 463, 499, 503–504, 506, 593
 Kelsey, J.L., 13, 137, 603
 Kendall, Sir M., 535, 604
 Kenny, S.D., 14, 604
 Kent, J.T., 471–472, 475, 530, 574, 576–577, 604
 Khurshid, A., 13, 612
 Kimball, A.W., 63, 604
 Klein, J.P., 606
 Kleinbaum, D.G., 137, 227, 604
 Knowler, W.C., 7, 14, 602
 Koch, G.G., 136, 145, 152, 189, 197, 235, 239, 283, 288, 445, 602, 604, 606, 613
 Korn, E.L., 207, 333, 471, 569, 604
 Koziol, J., 154, 602
 Kraemer, H.C., 235–236, 594, 604
 Kruskal W.H., 604
 Kruskal, W.H., 73, 376, 601
 Kudo, A., 172, 604
 Kupper, L.L., 55, 86, 137, 226–227, 603–604, 606
 Lachin, J.M., 38, 85, 91, 94, 100–101, 141, 164, 172–173, 175, 188, 223, 225–226, 238, 343, 348–351, 439, 460, 477, 484–486, 489–490, 495, 497–498, 504, 531, 594, 603–606, 610, 613, 615–616
 Lagakos, S.W., 439, 452, 454, 463, 493–494, 605
 Laird, N.M., 177, 180, 359, 402, 597, 605, 607
 Lakatos, E., 486, 605
 Lambert, D., 407, 605
 Lan, K.K.G., 504, 605–606
 Lan, S.L., 439, 490, 605–606
 Lancaster, H.O., 63, 606
 Landis, J.R., 189, 235, 239, 600, 606, 615
 Larsen, M., 7, 597
 Larsen, M.D., 321–322, 325, 603
 Laupacis, A., 56, 606
 Lavori, P.W., 5, 325, 350, 486, 488, 594, 603
 Lawless, J.F., 402, 413, 429, 597, 606
 Le, C.T., 85–86, 608
 Lee, E.T., 429, 606
 Lee, E.W., 499, 606
 Lee, P., 496, 610
 Lee, S., 520, 523, 600

- Lehmann, E.L., 64, 102, 108, 535, 539, 606
 Lemeshow, S., 233, 283, 302, 304, 345–346, 603
 Leung, H.K., 55, 606
 Levin, B., 13, 236, 239, 600
 Levin, M.L., 53, 606
 Lewis, D.K., 227, 604
 Lewis, E.J., 490, 606
 Lewis, J., 258, 616
 Liang, K.Y., 343, 588, 590, 598, 606, 616
 Limm L.L.-Y., 605
 Limm L.L.-Y., 493–494
 Lin, D.Y., 345, 463–464, 470, 474, 479, 498, 511, 513, 579, 607, 615
 Lin, J.S., 512, 607
 Lipsitz, S.H., 180, 607
 Little, C.R., 492, 494, 527, 601
 Little, R.J.A., 365, 607
 Liu, W., 70–71, 607
 Louis, T.A., 5, 19, 594, 607
 Lu B., 364
 Lu, B., 607
 Lunn, M., 494, 607
 Lupus Nephritis Collaborative Study Group, 439, 490, 605–606
 Machin, D., 85, 607
 Mack, T.M., 222, 607
 MacMahon, B., 53, 607
 Madalla, G.S., 573, 607
 Magee, L., 573, 607
 Makuch, R.W., 96–97, 607
 Mallows, C.L., 307, 607
 Mann, H.B., 102, 608
 Mantel, N., 39, 44, 64, 120, 123, 126, 144, 154, 217, 247, 305, 313, 449–451, 453–454, 602, 608, 610
 Marcus, R., 70, 608
 Marek, P., 451–452, 611
 Martin, D. C., 407
 Martin, D.C., 608
 Marubini, E., 429, 444, 494, 527, 608
 Mauer, S.M., 8, 595, 613
 McCanless, I., 445, 604
 McCullagh, P., 357, 403, 413, 580, 585–586, 588, 608
 McDermott, M., 495, 600
 McHugh, R.B., 85–86, 215, 223, 596, 599, 608
 McKinlay, S.M., 135, 608
 McNeil, D., 494, 607
 McNemar, Q., 213, 605, 608
 McPherson, K., 453–454, 610
 Mehrotra, K.G., 452, 608
 Mehta, C., 35–36, 608
 Meier, P., 131, 141, 432–433, 594, 603, 608
 Meng, R.C., 100, 609
 Michalek, J.E., 452, 608
 Miettinen, O.S., 18, 43–44, 53, 127, 224, 226, 232, 609
 Mihalko, D., 452, 608
 Miller, A.J., 305–306, 609
 Miller, H.D., 382, 597
 Miller, J.P., 495, 600
 Mills, S.D., 520, 523, 600
 Mogensen, C.E., 82, 609
 Moon, J.H., 520, 523, 600
 Moon, T., 495, 600
 Moore, D.F., 402, 609
 Morgenstern, H., 137, 227, 604
 Mori, M., 492–493, 610
 Morris, C.N., 180, 609
 Moses, L.E., 161, 594
 Moss, A.J., 495, 600
 Nagelkerke, N.J.D., 336, 609
 Nee, J.C.M., 239, 600
 Nelder, J.A., 357, 403, 413, 579–580, 585, 588, 608–609
 Nelson, W., 435, 502, 609
 Newcombe, R.G., 18, 609
 Neyman, J., 565, 597, 603, 609
 Noda A., 604
 Noda, A., 235
 Noether, G.E., 102, 108, 609
 Oakes, D., 429, 495, 597, 600
 O'Brien, R.G., 102, 107, 609
 Odeh, R.E., 85, 609
 Odell, P.M., 496–497, 610
 Oliver, D., 402, 605
 O'Quigley, J., 471–472, 475, 530, 604, 610
 Orav, E.J., 180, 607
 Ostfeld, A.M., 142, 616
 Paik, M.C., 13, 236, 239, 600
 Palesch, Y.Y., 498, 610
 Palta, M., 489, 610
 Patel, N., 35–36, 608
 Patterson, H.D., 293, 599
 Pearson, E.S., 16, 596
 Pearson, K., 57, 610
 Pee, D., 305, 600
 Pepe, M.S., 492–494, 610
 Peritz, E., 70, 608
 Perlman, M.D., 172, 610
 Peterson, A.V., 434, 492, 507, 512, 522, 610–611
 Peterson, B., 358, 610
 Peto, J., 450–454, 610
 Peto, R., 182–183, 258, 445, 450–454, 467, 492, 495–496, 596, 610, 616
 Petrov, B.N., 593
 Pettigrew, H.M., 31, 611
 Pfeffer, R.I., 222, 607
 Piantadosi, S., 463, 601
 Pike, M.C., 24, 222, 453–454, 496, 603, 607, 610–611

- Pinkel, D., 520, 523, 600
 Pirart, J., 8, 611
 Pitman, E.J.G., 108, 611
 Polansky, M., 5, 594
 Poulsen, J.E., 7, 597
 Pregibon, D., 302, 304, 611
 Preisser J.S., 364
 Preisser, J.S., 607
 Prentice, R.L., 341, 429, 450–452, 459, 465–469,
 471, 473, 492, 507, 512, 529, 595, 603, 611
 Pugh, T.F., 53, 607
 Qaqish, B.F., 364, 607
 Qu, Y., 103–104, 616
 Radhakrishna, S., 120, 158–159, 611
 Raghavarao, D., 85, 597
 Rahardja, D., 103–104, 616
 Ralston, A., 599
 Ramlau-Hansen, H., 503–504, 611
 Rao, C.R., 160, 535, 539, 544, 548, 556, 565, 611
 Rasmussen, N.K., 504, 593
 Reiner, E., 471, 610
 Robbins, H., 180, 611
 Roberts, R.S., 56, 606
 Robins, J.N., 129, 218, 228, 243, 261, 493–494,
 605, 611–612
 Rochon, J.N., 322, 324, 612
 Rohde, R.D., 490, 606
 Rolnitzky, L.M., 495, 600
 Ross, S.M., 382, 612
 Rothman, K.J., 137, 612
 Royall, R.M., 574, 612
 Rubenstein, L.V., 485, 612
 Rubin, D.B., 365, 607
 Russell, K.E., 307, 614
 Sabai, C., 341, 595
 Sackett, D.L., 56, 596, 606
 Sahai, H., 13, 612
 Santner, T.J., 421, 485, 601, 612
 Santucci, A., 494, 612
 SAS, 602, 612–613
 Schaid, D.J., 104–105, 613
 Schemper, M., 348, 471–472, 475, 610, 612
 Schlesselman, J.J., 137, 225, 612
 Schoenfeld, D., 348, 450, 452, 462, 471, 485–486,
 489, 523, 605, 612
 Schrek, R., 206, 612
 Schroeder, L.R., 520, 523, 600
 Schuirmann, D.J., 46, 612
 Schuster, J.J., 85, 612
 Scrucca, L., 494, 612
 Selawry, O.S., 520, 523, 600
 Selvin, S., 13, 613
 Sen, P.K., 43–44, 613
 Serfling, R.J., 543, 613
 Shannon, C.E., 334, 613
 Shoukri, M.M., 239, 613
 Sídk, Z., 64, 112, 452, 602
 Simon, R.M., 96–97, 307, 333, 471, 569, 604,
 607, 614
 Slager, S.L., 104–105, 613
 Sleight, T., 258, 616
 Smith, J.A., 223, 596
 Smith, P.G., 453–454, 610
 Snedecor, G.W., 303, 613
 Snell, E.J., 597
 Sparling, Y.H., 497, 613
 Spearman C., 73, 613
 Spurr, C.L., 520, 523, 600
 Staquet, M.J., 610
 Stare, J., 472, 612
 Starmer, C.F., 43–44, 152, 235, 288, 602, 613
 Statistical Analysis System (SAS), 345, 377, 420,
 466, 499, 513
 Steffes, M.W., 8, 595, 613
 Stokes, M.E., 136, 145, 189, 283, 613
 Stone, M.C., 308, 598
 Storrs, R., 520, 523, 600
 Straus, D.J., 492, 494, 527, 601
 Stuart, A., 235, 535, 604, 613
 Suchindran, C., 364, 607
 Sun, J., 495, 613
 Sun, X., 235, 613
 Sylvester, R.J., 610
 Tan, C.C., 492, 494, 527, 601
 Tang, D.I., 173, 613
 Tang, Y., 349–350, 613
 Tarone, R.E., 156, 263, 452, 601, 613
 Taylor, H.M., 382, 603
 Thall, P.F., 307, 420, 460, 477, 531, 605, 613–614
 Therneau, T.M., 470, 495, 600, 614
 Thisted, R.A., 32, 180, 269, 271, 370, 614
 Thomas, D.C., 227, 614
 Thomas, D.G., 31–32, 36, 80, 611, 614
 Thompson, W.D., 13, 137, 603
 Tibshirani, R.J., 302, 304, 602
 Tocher, K.D., 44, 614
 Truett, J., 307, 614
 Tsiatis, A.A., 459, 614
 Tsokos, C.P., 503, 594
 Turnbull, B.W., 495, 614
 Tygstrup, N., 594
 U.S. Surgeon General, 6, 614
 University Group Diabetes Program, 524–525, 614
 Vaeth, M., 313, 614
 Vail, S.C., 420, 614
 Valsecchi, M.G., 429, 444, 493–494, 527, 608
 Van Elteren, P.H., 64, 614
 Vyjeyanthi S.P., 604
 Vyjeyanthi, S.P., 235
 Wacholder, S., 227, 614
 Wald, A., 57, 62, 146, 562, 614
 Wallenstein, S., 183, 486, 615–616

- Wallis, W.A., 73, 604
Walter S.D., 615
Walter, S.D., 31, 53–55, 232, 615
Ward, J.F., 445, 604
Ware, J.H., 154, 452, 602, 613
Wedderburn, R.W.M., 579, 585, 609, 615
Wei L.J., 607
Wei, L.J., 164, 172–173, 345, 463–464, 470, 474, 498–499, 511–513, 579, 605–607, 615
Weinberg, C.R., 227, 614
Weissfeld, L., 498, 513, 615
White, A.A., 235, 615
White, H., 574, 579, 615
Whitehead, A., 258, 359, 615
Whitehead, J., 258, 615
Whitney, D.R., 102, 608
Whittemore, A.S., 13, 137, 321, 324, 603, 615
Wieand, S., 463, 601
Wilcoxon, F., 64, 615
Wilf, H.S., 599
Wilks, S.S., 16, 615
Williams, B.J., 507, 611
Wilson, E.B., 18, 42, 615
Wittes, J.T., 183, 486, 504, 606, 615–616
Wolfson, M., 364, 607
Wolman, L.J., 520, 523, 600
Woolf, B., 16, 27, 616
Wright, S.P., 71, 616
Wu, D.H., 492, 494, 527, 601
Yang I., 607
Yang, I., 511
Yang, Z., 235, 613
Ying, Z., 511, 607
Younes, N., 495, 497, 613, 616
Yusuf, S., 182–183, 258, 596, 616
Zeger, S.L., 343, 588, 590, 598, 606, 616
Zelen, M., 154, 496, 599, 616
Zhao, Q., 495, 613
Zhao, X., 495, 613
Zhao, Y.D., 103–104, 616
Zing, Z., 470, 607
Zuckerman, D.M., 142, 616
Zweifel, J.R., 30–31, 601

Index

- 2×2 Table**, 19–20
See also Matched Pairs and Stratified Analysis of 2×2 Tables
Cochran's Test, 40, 79
Conditional Hypergeometric Likelihood, 29
Fisher-Irwin Exact Test, 33
Likelihood Ratio Test, 42
Mantel-Haenszel Test, 39
Marginal Unadjusted Analysis, 121
Measures of Association, 20
Product Binomial Likelihood, 28
Product Binomial
 MLEs and Their Asymptotic Distribution, 249
Unconditional Test, 38
Aalen-Gill Test Statistics, 504
See also Counting Process and Weighted Mantel-Haenszel Tests
 G^P Family of Tests, 506
Logrank Test, 506
Wilcoxon Test, 506
Absolute Risk, 3
Accelerated Failure-time Model, 516
 Exponential, 514
 Log-Logistic Model, 518
 Weibull Model, 517
Actuarial Lifetable, 443
Akaike's Information Criterion, 307
Allowing for Clinic Effects in a Randomized Trial, 341
Analysis of Covariance (ANCOVA), 138
Antagonism, 137
Applications of Maximum Likelihood and Efficient Scores, 247
Aranda-Ordaz Link, 496
Asymptotic Distribution of the Efficient Score and the MLE, 556
Asymptotic Relative Efficiency, 110, 163, 165
 Competing Tests, 163
 Radhakrishna Family, 165, 194
Stratified Versus Unstratified Analysis of Risk Differences, 112, 118
Asymptotically Unbiased Estimates, 30
 Odds Ratio, 31
 Relative Risk, 31
Attributable Risk, 52
See also Population Attributable Risk
Barnard's Exact Unconditional Test for 2×2 Tables, 34
Best Linear Unbiased Estimator (BLUE), 548
Binomial Distribution, 14
 Asymmetric Confidence Limits, 15
 Asymptotic Distribution, 15
 Case of Zero Events, 19, 77
 Clopper-Pearson Confidence Limits, 16
 Complimentary Log-Log Confidence Limits, 17
 Exact Confidence Limits, 16
 Large Sample Approximations, 14
 Large Sample Variance, 15
 Logit Confidence Limits, 16

- Maximum Likelihood Estimation, 247, 275
 Test-Based Confidence Limits, 18
 Wilson Confidence Limits, 18
 Binomial Regression Model, 293
See also Logit Model
 Complimentary Log-Log Link, 294, 371
 Family of Models, 293
 Generalized Linear Models, 294
 Log Link, 294, 371
 Log Link Score Test, 373
 Logit Model, 286
 Probit Link, 294, 371
 Biomedical Studies, Types of, 5
 Biostatistics, 2
 Bowker's Test of Symmetry
 Matched Data, 234, 246
 Breslow-Day Test of Homogeneity, 155
See also Tarone's Test
 $C(\alpha)$ Test, 565
See also Score Test
 Case-Control Study, 6, 201, 220
 Matched, 220
 Unmatched, 201
 Cauchy-Schwartz Inequality, 160, 556
 Causal Agent, 6
 Cause-Specific Hazard Function, 492
 Censoring
 At Random, 430–431
 Interval, 465, 495
 Right, 430
 Central Limit Theorem, 537
 Liapunov's Theorem, 539
 Lindberg-Levy Theorem, 539
 Multivariate Case, 540
 Clinical Trial, 6
 Clopper-Pearson Confidence Limits, 16
 Cochran-Armitage Test for Trend, 74, 84
 Stratified, 192
 Cochran-Mantel-Haenszel Test, 41, 126
 Correlation Test, 73
 General Association Test, 68
 Mean Scores Test, 63
 Cochran's Model for Stratified Versus Unstratified
 Analysis of Risk Differences, 140
 Cochran's Poisson Variance Test, 388
 Cochran's Stratified Test of Association
 2×2 Tables, 125, 184, 187
 As a $C(\alpha)$ Test, 261
 Pair-Matched Tables, 230
 Radhakrishna Family, 158
 Relative Risks of Poisson Intensities, 422
 Cochran's Test for 2×2 Table, 40, 257
 Cochran's Test of Homogeneity, 153
 Expected Value, 178
 Stratified 2×2 Tables, 153
 Stratified Analysis of Pair-Matched Tables, 230
 Stratified Relative Risks of Poisson Intensities, 422
 Cohort Study, 6
 Comparison of Survival Probabilities for Two Groups, 436
 Comparison of Weighted Tests for Stratified 2×2 Tables, 174
 Competing Risks, 492
 Cause-Specific Hazard Function, 492
 Crude (Mixed) Rate, 493
 Exponential Survival, 519
 Net (Pure) Rate, 493
 Subdistribution Function, 493
 Complementary Log-Log Transformation, 17
 Assessing the PH Model Assumption, 470
 In Discrete Time PH Model, 466
 Of a Probability, 76
 Of Survival Function, 435
 Conditional Generalized Linear Models for
 Matched Sets, 588
 Conditional Hypergeometric Likelihood, 29
 Maximum Likelihood Estimation, 257, 277, 281
 Score Test, 257
 Stratified, 274
 Conditional Independence, 209, 215
 Matched Pairs, 209
 Conditional Large Sample Test and Confidence
 Limits for Conditional Odds Ratio, 217
 Conditional Large Sample Variance for 2×2 Table, 40, 80
 Conditional Logistic Regression Model for
 Matched Sets, 337
 Clinic Effects, 341
 Explained Variation, 348
 Fitting the Model, 341
 Madalla's R^2 , 348
 Maximum Likelihood Estimation, 339
 PROC PHREG, 341
 Retrospective Study, 340
 Robust Inference, 345
 Conditional Mantel-Haenszel Analysis for
 Matched Pairs, 260
 Conditional Odds Ratio for Matched Pairs, 215
 Case-Control Study, 221
 Conditional Large Sample Test and Confidence
 Limits, 217
 Exact Confidence Limits, 214
 Large Sample Confidence Limits, 215
 Large Sample Variance, 216
 Retrospective, 221
 Stratified Tests of Association and
 Homogeneity, 229
 Conditional Poisson Regression Model for
 Matched Sets, 410
 Conditional Within-Strata Analysis, 121
 Confounding, 137

- Confounding and Effect Modification, 137
- Consistent Estimator, 538
 - \sqrt{n} -Consistent, 539
- Contingency Chi-Square Test, 38
 - See also* Pearson Chi-Square Test
- Convergence in Distribution
 - Slutsky's Theorem, 542
 - Transformations, 544
- Convergence in Probability, 536
 - Slutsky's Theorem, 544
- Count Data, 381
- Counting Process, 500
 - Aalen-Gill Test Statistics, 504
 - Cumulative Intensity, 501
 - Filtration, 501
 - Intensity, 501
 - Intensity Estimate, 503
 - Kernel-smoothed Intensity Estimate, 503
 - Martingale, 501
 - Martingale Transform, 505
 - Martingale
 - Compensator, 502
 - Submartingale, 502
 - Nelson-Aalen Estimate of Cumulative Intensity, 502
 - Predictable Process, 505
 - Stochastic Integral, 505
- Cox's Logit Model for Matched Pairs, 215
- Cramér-Rao Inequality, 555
 - Efficient Estimates, 555
- Cross-Sectional Study, 5, 14
- Cumulative Hazard Function, 430
 - Kaplan-Meier Estimate, 435
 - Nelson-Aalen Estimate, 435
- Cumulative Incidence, 10
- Cumulative Intensity, 382
- Cumulative Intensity Function
 - Nelson-Aalen Estimate, 502
- δ -Method, 541
 - Multivariate Case, 542
- DerSimonian and Laird Random Effects Model for Stratified 2×2 Tables, 177, 197
- Deviance, 315, 580, 584
- Diabetes, 3
 - Diabetes Control and Complications Trial, 9, 219, 231, 245, 298, 313, 315, 319, 330, 332, 337, 342, 361, 364, 386, 388, 392, 395, 401, 404, 406–407, 413–414, 445, 455, 481–482, 506, 509, 512
- Diabetes
 - Hypoglycemia, 11
 - Nephropathy, 4
 - Albumin Excretion Rate, 4
 - Microalbuminuria, 4, 10–11, 82
 - Natural History, 7
 - Neuropathy, 22, 32
- Retinopathy, 4
- Direct Adjustment Using Logistic Regression, 373
- Discrete or Grouped Time Lifetable, 443
- Doubly Homogeneous Poisson Model, 382
- Effect Modification, 138
- Efficiency, 108, 159–160
 - Cramér-Rao Inequality, 555
 - Estimation Efficiency, 111
 - Pitman Efficiency, 108
- Efficient Score, 552
 - Asymptotic Distribution, 556
- Efficient Score Test, *See* Score Test
- Efficient Tests, 111
 - Radhakrishna Family for Stratified 2×2 Tables, 158
 - Risk Difference for Stratified 2×2 Tables, 114
 - Stratified Conditional Odds Ratio for Matched Pairs, 230
 - Stratified Marginal Relative Risk for Matched Pairs, 230
- Entropy R^2
 - 2×2 Table, 49
 - Logistic Regression Model, 334
- Entropy Loss in Logistic Regression Model, 334
- Epanechnikov's Kernel, 504
- Epidemiology, 2
 - Population-Based, 3
- Equivalence and Noninferiority, 45
- Equivalence
 - Odds Ratio, 46
 - Relative Risk, 46
 - Risk Difference, 46
 - Two One-Sided Tests (TOST), 46
- Estimation Efficiency, 111
- Estimation Precision and Sample Size, 86
- Event Rate, 11, 382
- Event-Time Data, 429
- Exact Confidence Limits
 - A Probability, 16
 - Conditional Odds Ratio for Matched Pairs, 215
 - Odds Ratio for Independent Groups, 32
 - Relative Risk for Independent Groups, 33, 80
 - Risk Difference for Independent Groups, 33, 80
- Exact Inference for 2×2 Table, 32
- Exact Test
 - Barnard's Unconditional Test for 2×2 Table, 34
 - Fisher-Irwin Test for 2×2 Table, 33
 - Matched Pairs, 211
 - Polychotomous and Ordinal Data, 76
- Example
 - A 3×3 Table (Sample Size), 102
 - A Hypothetical Example (Kappa Index of Agreement), 237
- Actuarial Lifetable in PROC LIFETEST, 447
- Albuminuria in the DCCT (Sample Size), 103
- ARE of Normal Median:Mean, 111

- Case-Control Study (Matched Sample Size), 226
- Cholesterol and CHD (Number Needed to Treat), 56
- Clinical Trial in Duodenal Ulcers (Stratified 2×2 Tables), 122, 126, 129, 132, 141, 149, 153, 162, 168, 171, 174, 181, 293
- Conditional MLE, Ulcer Clinical Trial, 274
- Conditional Power (McNemar's Test), 225
- Coronary Heart Disease in the Framingham Study, 73
- Cochran-Armitage Test for Trend, 76
 - Interaction in Logistic Regression, 327
 - Logit Model, 287, 290
 - Population Attributable Risk, 55
 - Sample Size, 106
- Correlated Observations (Weighted Least Squares), 550
- Covariate Adjusted Survival Estimates, 480
- DCCT Hypoglycemia Incidence (Recurrent Events), 506, 509, 512
- DCCT Nephropathy Data
- GEE Repeated Measures, 364
 - HbA_1c by Blood Pressure Interaction, 332
 - Logistic Model Robust Variance, 319
 - Logistic Model Score Test, 313,
 - Logistic Model, 337
 - PROC GENMOD, 315
 - Treatment by Duration Interaction, 330
- DCCT Time-Dependent HbA_1c and Nephropathy, 481
- DCCT Treatment Group Adjusted for Clinic Logistic Mixed Model, 361
- Logistic Model, 342
- Estrogen Use and Endometrial Cancer (Matched Case-Control Study), 222
- Ethnicity and Hypertension in the Diabetes Prevention Program (Multiple Tests), 71
- Ethnicity and Hypertension in the Diabetes Prevention Program (Stratified $R \times C$ Tables), 189
- Exact Inference, 35
- Exact Inference Data, 41, 44
- Frequency Matching, 207
- Genetic Marker (Sample Size), 105
- Heteroscedasticity (Weighted Least Squares), 549
- Homoscedastic Normal Errors Regression Information Sandwich, 578
- MLE from Non-*i.i.d.* Observations, 560
- Hospital Mortality
- A Proportion, 18
 - Poisson MLE, 560
- Hypoglycemia in the DCCT
- Negative Binomial Model, 413–414
 - Poisson Regression, 395, 404, 406
- Rates, 386, 388, 392, 401
- Zeros-inflated Poisson Model, 407
- Hypothetical *p*-values (Multiple Tests), 73
- Hypothetical Data (Conditional Logit Model for Matched Pairs), 268
- Ischemic Heart Disease (Logistic Regression in Unmatched Retrospective Study), 308
- Large Sample (Matched Pairs), 213, 217
- Log Odds Ratio, 259
- Log(*p*)
- δ -Method, 542
 - Slutsky's Theorem, 545
- Low Birth Weight, 233
- Conditional Logistic Model, 345
- Lupus Nephritis Study (Survival Sample Size), 490
- Member-Stratified Matched Analysis, 232
- Meta-analysis of Effects of Diuretics on Pre-eclampsia, 182
- Multinomial Distribution (Central Limit Theorem), 540
- Multinomial Generalized Logits
- Multivariate δ -Method, 543
 - Slutsky's Theorem, 546
- Multiple Linear Regression Model (Explained Variation), 571
- National Cooperative Gallstone Study: Biopsy Substudy
- Multinomial Logistic Model, 354
 - Proportional Odds Model, 358
- Nephropathy in the DCCT (Lifetables), 445, 455
- Neuropathy Clinical Trial (2×2 Table), 22–24, 26–27, 32, 41, 43–44, 47
- Normal Errors Model Score Test, 568
- Planning a Study (Sample Size for), 94
- Planning a Study for Equivalence or Noninferiority (Sample Size for), 98
- Poisson-Distributed Counts
- Information Sandwich, 578
 - Maximum Likelihood Estimation, 559
 - Tests of Significance, 568
- Pregnancy and Retinopathy Progression, 219, 231
- Recombination Fraction
- Generalized Logits, 61
 - Newton-Raphson, 271
 - Pearson Goodness-of-Fit Test, 59
 - Sample Size, 100
- Religion and Mortality (Stratified 2×2 Tables), 142, 149, 154, 163, 169, 171, 174, 182
- Retinopathy in the DCCT
- Rank Tests, 65
 - Stratified Mean Scores Test, 191
 - Weighted Kappa, 238
- Robust Information Sandwich, 477

- Sample Size for a Nested Case-control Study (Conditional Logistic Model), 351
- Simple Proportion (Central Limit Theorem), 539
- Simpson's Paradox, 144
- Single Proportion (Sample Size For), 96
- Small Sample
- Exact Limits for Matched Pairs, 216
 - Exact Test for Matched Pairs, 212
- Smoking and Lung Cancer (Case-Control Study), 206
- Squamous Cell Carcinoma (Survival Analysis), 439, 454, 474
- Stratified Analysis (Sample Size for Logistic Regression), 323
- Study of Post-traumatic Stress Disorder (Sample Size for Logistic Regression), 322, 325
- Test of Proportions (Sample Size for Logistic Regression), 325
- Testing the Proportional Hazards Assumption, 479
- The 2×2 Table (Pearson Chi-Square Test), 58
- Three Strata with Heterogeneity (Power and Sample Size), 187
- Two Homogeneous Strata (Radhakrishna Family), 164, 168
- Two Strata (ARE Versus Unstratified), 115
- Ulcer Clinical Trial
- Stratified 2×2 Tables, 254, 260–261
 - Stratum By Group Interaction, 329
- Unconditional Sample Size (McNemar's Test), 224
- Explained Variation, 569
- Conditional Logistic Model, 348
- Entropy R^2 , 49
- Entropy Loss, 573
- Logistic Regression Model, 333, 376
- Madalla's R^2_{LR} , 573
- Negative Log-likelihood Loss, 573
- PH Model, 471
- Poisson Regression Model, 401, 425
- Residual Variation, 572
- Squared Error Loss, 570
- Uncertainty Coefficient, 49
- Exponential Survival Distribution, 430, 483, 513
- Accelerated Failure-time Model, 514
 - Maximum Likelihood Estimation, 513
- Family of Binomial Distribution Regression Models, 293
- Family of Tests, 163
- G^p Family of Tests for Hazard Functions, 452
 - Radhakrishna Family for 2×2 Stratified Tables, 158, 165
 - Weighted Mantel-Haenszel Tests, 449
- First-Step Iterative Estimate, 179, 391
- Fisher Scoring, 259, 270
- Fisher-Irwin Exact Test, 33
- Fisher's Information Function, 553
- See also* Information
- Fixed Effects, 185
- Fixed-Point Iterative Method, 180, 391
- Frailty Model, 463
- Frequency Matching, 207
- G^p Family of Tests for Survival Functions, 452
- See also* Weighted Mantel-Haenszel Test
- Gallstones, 81
- Gastwirth Maximin-efficient Robust Test (MERT), 170
- G^p Family, 453
 - Radhakrishna Family, 170
 - Scale Robust Test, 170
- Gauss-Markov Theorem, 548
- Gehan-Wilcoxon Test for Lifetables, 451
- Generalized Additive Models, 302, 304
- Generalized Estimating Equations (GEE), 588
- Information Sandwich Variance Estimate, 591
 - Logistic Multivariate Model, 368
 - Logistic Repeated Measures Model, 364
 - Multiple Count Outcomes, 419
 - Poisson Regression Model, 404
 - Score Test, 591
 - Working Correlation, 589
 - Exchangeable, 590
 - First-order Autoregressive, 590
 - Independence, 589
 - Unstructured, 590
- Generalized Linear Models, 579
- Binomial Regression Models, 293
 - Canonical Link Function, 583
 - Chi-Square Goodness of Fit, 585
 - Conditional for Matched Sets, 588
 - Deviance, 584
 - Exponential Family, 581
 - Generalized Estimating Equations (GEE), 588
 - Link Function, 580
 - Minimum Chi-Square Estimation, 586
 - Quasi-likelihood Functions, 586
 - SAS PROC GENMOD, 289
- Greenwood's Estimate of Variance of Survival Function Estimate, 435
- Haldane-Anscombe Estimates, 31
- Hazard Function, 430
- Cause-Specific Competing Risk, 492
 - Estimation, 435
 - Kaplan-Meier Estimate, 521
- Hessian, 554
- Heterogeneity, 138, 150
- Homogeneity, 139
- Homogeneous Poisson Process, 382
- Homogeneous Poisson Regression Model, 393
- Hypergeometric Distribution, 28
- Central, 34, 79–80

- Large Sample Approximation, 40
- Noncentral, 29, 130
- Incidence, 11, 14
- Information, 554
- Information Sandwich Variance Estimate, 574
 - Generalized Estimating Equations (GEE), 591
 - Logistic Regression Model, 319, 374
 - Poisson Regression Model, 404
 - Proportional Hazards Models, 463
 - Robust Score Test in Logistic Regression Model, 319
 - Wald Test in Logistic Regression Model, 319
- Information
 - Estimated Information, 558
 - Expected Information, 554
 - Information Equality, 554
 - Information Function, 554
 - Observed Information, 554
- Intensity
 - Counting Process, 501
 - Poisson Process, 382
- Intent-to-Treat
 - Lifetable, 494
 - Principle, 3
- Interactions, 138, 150
 - Logistic Regression Model, 325, 375
 - Qualitative-Qualitative Covariate Interaction, 326
 - Quantitative Covariate Interaction, 330
 - PH Regression Model, 460
- Interval Censoring, 443, 465, 495
- Intraclass Correlation, 239
- Invariance Principle (of MLE), 558
- Invariance Under Transformations, 558
- Iterative Maximum Likelihood, 269
- Iteratively Reweighted Least Squares (IRLS), 550
- Kalbfleisch-Prentice Marginal PH Model, 467
- Kaplan-Meier Estimate
 - Cumulative Hazard Function, 435
 - Hazard Function, 522
 - Survival Function, 432
- Kappa Index of Agreement
 - Duplicate Binary Gradings, 235
 - Duplicate Polychotomous or Ordinal Gradings, 237
 - Intraclass Correlation, 239
 - Large Sample Variance, 236
 - Weighted Kappa, 238
- Kernel-smoothed Intensity Estimate, 503
- Kruskal-Wallis Test, 73
- Law of Large Numbers (Weak Law), 537
- Least Squares Estimation, 546
 - Gauss-Markov Theorem, 548
 - Iteratively Reweighted Least Squares, 550
 - Ordinary Least Squares, 546
 - Weighted Least Squares, 548
- Least Squares Means (LSMEANS)
 - SAS
 - PROC GENMOD, 314
 - Liapunov's Central Limit Theorem, 539
 - Lifetable Construction, 441
 - Likelihood Function, 551
 - Likelihood Ratio Test, 563
 - $R \times C$ Table, 42
 - 2×2 Table, 42, 255
 - Composite Test, 563
 - Conditional Logit Model (Matched Pairs), 268
 - Logistic Regression Model, 309
 - Logit Model, 255
 - Matched Pairs, 268
 - Proportional Hazards Models, 462
 - Test of a Subhypothesis, 563
 - Type III Option in PROC GENMOD, 310, 315
- Lindberg-Levy Central Limit Theorem, 539
- Linearized Rate, 383
- Link Function, 302
- Lin's Test of the PH Assumption, 470
- Local Alternative, 108, 159, 194
- Log Risk Model
 - Maximum Likelihood Estimation, 277
- Logistic Function, 16–17, 82
- Logistic Model
 - Cox's Adjustment for Ties in the PH Model, 467, 528
- Logistic Regression and Binomial Logit Regression, 286
- Logistic Regression Model, 283
 - See also* Conditional Logistic Regression Model for Matched Sets
 - Conditional Model for Matched Sets, 339
 - Confidence Limits on Conditional Probability, 286
 - Direct Adjustment, 373
 - Disproportionate Sampling, 307
 - Entropy Loss, 334
 - Explained Variation, 333, 376
 - GEE Multivariate Model, 368
 - GEE Repeated Measures Model, 364
 - Independent Observations, 283
 - Information Sandwich Variance Estimate, 319, 374
 - Interactions, 325, 375
 - Qualitative-Qualitative Covariate Interaction, 326
 - Quantitative Covariate Interaction, 330
 - Interpretation, 294
 - Likelihood Ratio Test, 309
 - Model Test, 309
 - Test of Model Components, 309
 - Log(X) Transformation, 371
 - Madalla's R^2 , 336
 - Max Rescaled R^2 , 337

- Maximum Likelihood Estimation, 283, 370
 Model Coefficients and Odds Ratios, 294
 Models for Polychotomous or Ordinal Data, 352
 Multinomial, 353
 Multivariate or Repeated Measures, 363
 Newton-Raphson Iteration, 287
 Overdispersion, 317
 Partial Regression Coefficients, 302
 Power and Sample Size, 321, 374
 Adjusted for Multiple Factors, 324
 Multiple Categorical Effects, 322
 Proportional Odds Model, 357
 Random Coefficient Model, 369
 Random Effects and Mixed Models, 359
 Random Intercept Model, 359
 Random Treatment Effect, 361
 Residual Variation Explained, 335
 Robust Inference, 317
 SAS Procedures, 288
 Score Test, 310, 373
 Model Test, 310
 Test of Model Components, 312
 Squared Error Loss, 333
 Stepwise Procedures, 304
 Stratified 2×2 Tables, 291
 Unconditional Model for Matched Sets, 338
 Unmatched Case-Control Study, 308
 Wald Tests, 312
 Logit Confidence Limits
 Probability, 16
 Survival Function, 522
 Logit Model, 82, 250
 2×2 Table, 82
 Binomial Regression Model, 286
 Matched Case-Control Study, 269
 Matched Pairs
 Conditionally, 265, 280
 Unconditionally, 263
 Maximum Likelihood Estimation, 250, 276
 Logits, 60
 $R \times C$ Table, 69
 Continuing Ratio, 60
 Cumulative, 61
 Generalized, 60, 66
 Pairwise, 60
 Log-Logistic Survival Distribution, 518
 Accelerated Failure-time Model, 518
 Logrank Test, 450
 Aalen-Gill Family Test, 506
 As a PH Regression Model Score Test, 528
 Weighted Mantel-Haenszel Test, 450
 Madalla's R^2_{LR} , 573
 Conditional Logistic Model, 348
 Logistic Regression Model, 336
 PH Regression Model, 472
 Poisson Regression Model, 401
 Mallow's C_p , 307
 Mantel-Haenszel Analysis, 121
 Matched Pairs, 217, 242
 Multiple Matching, 232
 Pair-Matched Tables, 228
 Stratified 2×2 Tables, 121
 Mantel-Haenszel Estimates, 126, 194
 Matched Pairs, 217
 Stratified 2×2 Tables, 126
 Stratified-Adjusted Odds Ratio, 127
 Large Sample Variance of Log Odds Ratio, 128
 Stratified-Adjusted Relative Risk, 127
 Mantel-Haenszel Test, 40, 125
 2×2 Table, 40
 Matched Pairs, 217
 Null and Alternative Hypotheses, 126, 150
 Power and Sample Size, 183
 Score Test for 2×2 Table, 258
 Score Test for Stratified 2×2 Tables, 260
 Stratified 2×2 Tables, 123
 Weighted, for Lifetables, 449
 Mantel-Logrank Test, 450
 See also Cochran-Mantel-Haenszel Test
 As PH Model Score Test, 528
 Marginal Relative Risk for Matched Pairs
 Prospective Study, 218
 Retrospective Study, 222
 Stratified Analysis, 230
 Martingale, 502
 See also Counting Process
 Matched Case-Control Study, 220
 Conditional Logit Model, 268
 Conditional Odds Ratio, 220, 269
 Matched Pairs, 208
 Case-Control Study, 220
 Conditional Logit Model, 216, 265, 280
 Conditional Odds Ratio, 215, 223
 Correlation, 241
 Cross-Sectional Or Prospective, 208
 Exact Test, 211
 Kappa Index of Agreement, 235
 Mantel-Haenszel Analysis, 217
 Marginal Relative Risk, 218, 222
 McNemar's Test, 212
 Measures of Association, 214
 Stratified Analysis, 227
 Tests of Marginal Homogeneity, 211
 Tests of Symmetry, 211
 Unconditional Logit Model, 263
 Matched Polychotomous Data, 234
 Bowker's Test, 234
 Marginal Homogeneity, 235
 McNemar's Test, 234
 Quasi-symmetry Homogeneity, 235
 Matching, 6, 206, 215, 220

- Matching Efficiency, 226
 Matrices, 536
 Maximin Efficiency, 169
 Maximin-efficient Robust Test (MERT)
 G^p Family, 453
 Gastwirth Scale Robust MERT, 170
 Radhakrishna Family, 170
 Stratified Pair Matching, 231
 Wei-Lachin Test of Stochastic Ordering, 171
 Maximum Likelihood Estimation, 551
 Asymptotic Distribution of MLE, 556
 Asymptotic Distribution of Score, 556
 Binomial Distribution, 276
 Conditional Hypergeometric Likelihood, 257,
 277, 281
 Conditional Logistic Model, 339
 Consistency and Asymptotic Efficiency of the
 MLE, 557
 Efficient Score, 552
 Estimated Information, 558
 Estimating Equation, 551
 Expected Information, 554
 Exponential Survival Distribution, 513
 Fisher Scoring, 270
 Independent But Not Identically Distributed
 Observations, 560
 Information, 554
 Information Inequality, 554
 Invariance Under Transformations, 558
 Likelihood Function, 551
 Log Risk Model, 277
 Logistic Regression Model, 284, 370
 Logit Model, 276
 Logit Model for 2×2 Table, 250
 Multiplicative Intensity Model, 508
 Newton-Raphson Iteration, 269
 Observed Information, 554
 Odds Ratio for Independent Groups, 30
 Odds Ratio in Stratified Product Binomial
 Likelihood, 279, 282
 Poisson Model, 383
 Poisson Regression Model, 394
 Proportional Hazards Model, 461
 Relative Risk in Stratified Product Binomial
 Likelihood, 280
 Stratified Conditional Hypergeometric
 Likelihood, 274, 278
 Stratified-Adjusted Odds Ratio, 130
 Tests of Significance, 560
 Weibull Survival Distribution, 515
 McNemar's Test, 212, 217–218, 234, 242–243,
 268, 281
 Mean Square Error
 Variance and Bias, 134
 Measurement Error Model, 175, 196
 Measures of Association
- $R \times C$ Table|069 Measures of Association
 2×2 Table, 20
 Matched Pairs, 214
 Measures of Relative Risk in 2×2 Table, 20
 Meta-analysis, 120, 177, 182
 Minimum Chi-Square Estimation, 586
 Minimum Variance Linear Estimates (MVLE)
 Joint Stratified-Adjusted Estimates, 189
 Pair-Matched Tables, 229
 Stratified-Adjusted, 131, 193, 422
 Versus Mantel Haenszel Estimates, 134, 193
 Minimum Variance Unbiased Estimates (MVUE),
 556
 Missing at Random (MAR)
 Missing Completely at Random (MCAR), 589
 Model Building: Stepwise Procedures, 304
 Backwards Elimination, 305
 Cross-Validation, 307
 Forward Selection, 305
 Reduced Model Bias, 305
 Modified Kaplan-Meier Estimate, 444
 Moment Estimate, 197
 Measurement Error Model, 197
 Random Effect Variance Component, 179
 Poisson Model, 390
 Recombination Fraction, 272
 Multinomial Distribution, 42, 56
 Central Limit Theorem, 540
 Large Sample Approximation, 57, 83
 Wald Test, 57, 83
 Multinomial Logistic Regression Models, 353
 Multiple Matching
 Mantel-Haenszel Analysis, 232
 Multiple Tests, 69
 Bonferroni Adjustment, 70
 Closed Test Procedures, 71
 Hochberg Procedure, 71
 Holm Procedure, 71
 Hommel Procedure, 71
 Multiplicative Intensity Model, 507
 Likelihood Function, 508
 Maximum Likelihood Estimation, 508
 Multivariate Null Hypothesis, 145
 Multivariate or Repeated Measures Models, 363
 Multivariate Tests of Hypotheses, 145
 Natural History of Disease Progression, 3
 Nature of Covariate Adjustment, 136
 Negative Binomial Model, 412
 Negative Binomial Distribution, 412
 Negative Binomial Regression Model, 414
 Power and Sample Size, 418
 Negative Log-likelihood Loss, 573
 Poisson Regression Model, 401
 Nelson-Aalen Estimate
 Cumulative Hazard Function, 435
 Cumulative Intensity Function, 502

- Hazard Function, 435
- Survival Function, 436
- Newton-Raphson Iteration, 269
- Neyman-Pearson Hypothesis Test, 36
 - General Considerations, 36
- NIH Model, 177
- Noncentrality Factor, 91, 99
- Noncentrality Parameter
 - 0.03 Noncentrality Parameter, 91
 - 1.0 Noncentrality Parameter, 99, 109
 - Noninferiority
 - Odds Ratio, 47
 - Relative Risk, 47
 - Risk Difference, 46–47
 - Notation, 535
 - Number Needed to Treat, 56, 82
 - Odds Ratio, 26
 - 2 × C Table, 63
 - Asymptotic Distribution, 26
 - Conditional (Matched Pairs), 215
 - Log Odds Ratio
 - Asymptotic Distribution, 26
 - Large Sample Variance, 27, 78
 - Logistic Regression Model Coefficients, 295
 - Retrospective, 202
 - Omnibus Test, 146
 - MANOVA Repeated Measures Test, 367
 - Null and Alternative Hypotheses, 146
 - Partitioning of the Alternative Hypothesis, 149
 - Stratified 2 × 2 Tables, 148
 - Optimal Weights
 - Efficient Tests for Stratified 2 × 2 Tables, 160
 - MVLE, 132
 - Weighted Mantel-Haenszel Test for Lifetables, 450
 - Weights Inversely Proportional to the Variances, 114, 118, 131, 134, 158
 - Ordinary Least Squares (OLS), 546
 - Overdispersion
 - Logistic Regression Model, 317
 - Poisson Model, 388
 - Poisson Regression Model, 402
 - Stratified 2 × 2 Tables, 178
 - Pair and Member Stratification for Matched Pairs, 227
 - Pair-Matched Retrospective Study, 220
 - See also* Matched Case-Control Study
 - Partial Association, 119
 - Partial Correlation, 140
 - Partitioning of the Omnibus Null and Alternative Hypotheses, 149
 - Partitioning of Variation, 134, 178, 196, 240, 537
 - Pearson Chi-Square Goodness of Fit, 585
 - Pearson Chi-Square Test
 - $R \times C$ Table, 38, 67
 - 2 × 2 Table, 38
 - 2 × C Table, 62
 - Equivalence to Z -Test for Two Proportions, 80
 - Goodness-of-Fit Test, 59
 - Generalized Linear Models, 585
 - Poisson Regression Model, 403
 - Homogeneity of Matched Sets, 230
 - One Sample, 57
 - Partitioning, 63
 - Power and Sample Size, 100
 - Pearson Goodness of Fit, 402
 - Peto-Breslow Adjustment for Ties in the PH Model, 467
 - Peto-Peto-Prentice-Wilcoxon Test for Lifetables, 451, 524
 - Pitman Efficiency, 110, 159
 - Poisson Distribution, 382
 - Poisson Model
 - Cochran's Variance Test, 388
 - Doubly Homogeneous, 382
 - Information Sandwich Variance Estimate, 420
 - Maximum Likelihood Estimation, 383
 - Overdispersion, 389
 - Overdispersion Variance Component Estimation, 420
 - Random Effects Model, 389
 - Stratified MVLE of Relative Risks, 422
 - Poisson Process, 382
 - Cumulative Intensity, 382
 - Homogeneous, 382
 - Intensity, 382
 - Poisson Regression Model, 393
 - Applications, 401
 - Conditional Model for Matched Sets, 410, 426
 - Explained Variation, 401, 425
 - Information Sandwich Variance Estimate, 404
 - Madalla's R^2 , 401
 - Maximum Likelihood Estimation, 394
 - Multiple Outcomes, 419
 - Negative Log-likelihood Loss, 401
 - Overdispersion, 402
 - Pearson Goodness-of-Fit Test, 403
 - Power and Sample Size, 416
 - Quasilikelihood Estimation, 402
 - Robust Inference, 404
 - Score Test, 425
 - Squared Error Loss, 401
 - Unconditional Model for Matched Sets, 410
 - Zeros-Inflated Model, 407
 - Polychotomous and Ordinal Data, 57
 - Conditional Models for Matched Sets, 359
 - Logistic Regression Models, 352
 - Multinomial Logistic Regression Models, 353
 - Multiple Independent Groups, 67
 - One Sample, 57
 - Proportional Odds Model, 357
 - Stratified $R \times C$ Tables, 188

- Two Samples, 62
Polychotomous Data
 Matched, 234
Population Attributable Risk, 53
 Asymptotic Distribution, 54
 Large Sample Variance of Logit, 54, 81
 Matched Pairs, 219
 Retrospective Study, 205, 240
Population-Averaged Odds Ratio, 214
Population-Averaged Relative Risk, 218
Power and Sample Size, 87
 Z-Test
 General, 87
 Means in Two Groups, 116
 Poisson Intensities in Two Groups, 117, 423
 Proportions in Two Groups, 92, 116,
Chi-Square Tests, 99
 Cochran-Armitage Test of Trend, 104
 Cochran's Test of Association, 186
 Conditional Logistic Regression Model, 348
 Binary Covariate, 351
 Multivariate Model, 351
 Quantitative Covariate, 350
Cox's PH Model, 486
 Adjusted Analysis, 489
 Qualitative Covariate, 487
 Quantitative Covariate, 488
Equivalence
 Proportions in Two Groups, 96, 118
Logistic Regression Model, 321, 374
 Adjusted for Multiple Factors, 324
 Multiple Categorical Effects, 322
Logrank Test, 483, 529
 Maximum Information Design, 489
McNemar's Test, 223, 243
 Conditional, 224
 Unconditional, 223
Mean Score (Rank) Test, 102
Negative Binomial Model, 418
Noninferiority
 Proportions in Two Groups, 98, 118
Pearson Chi-Square Test, 100
 2×2 Table, 116
 Multinomial Single Sample, 100
Poisson Regression Model, 416
Radhakrishna Family of Tests of Association,
 184
Simplifications, 91
Survival Analysis, 483
Test for Exponential Hazards, 483, 529
The Fundamental Relationship, 90
Wald Test in Poisson Regression Model, 423
 Wilcoxon and Mann-Whitney Tests, 102
Power Function, 88
Precision, 85
Prevalence, 14
Probability as a Measure of Risk, 14
Probit Regression Model, 371
 See also Binomial Regression Model, Probit
 Link
Product Binomial Likelihood
 2×2 Table, 28
 Logit Model, 250
 Maximum Likelihood Estimation, 250
 Stratified, 262, 279
Product-Limit Estimator, *See* Kaplan-Meier
 Estimate
Profile Likelihood, 256
Proportional Hazards Alternative, 451
Proportional Hazards Models, 456
 See also Multiplicative Intensity Model
Adjustments for Ties, 464
 Cox's Logistic Model, 466, 528
 Kalbfleisch-Prentice Marginal Model, 467
 Maximum Likelihood Estimation for the
 Peto-Breslow Likelihood, 527
 Peto-Breslow Approximate Likelihood, 467,
 527
 Prentice-Gloeckler Model, 465
Discrete and Grouped Data, 465
Discrete Time, 465
Explained Variation, 471
 Kent-O'Quigley Measures, 471
 Madalla's R^2 , 472
 Schemper's V_2 , 471
Fitting the Model, 461
Frailty Model, 463
Full Likelihood Function, 459
Information Sandwich Variance Estimate, 463
Likelihood Ratio Tests, 462
Maximum Likelihood Estimation, 461
Partial Likelihood Function, 457
PH Model Assumptions, 469
 Cox's Test, 469
 Lin's Test, 470
 Log-Log Survival Plots, 470
Robust Inference, 463
Robust Score Test, 463
Robust Wald Test, 463
Schoenfeld (Score) Residual, 462
Score Test in the Peto-Breslow Likelihood, 528
Score Tests, 462
Stratified Models, 460
Survival Function Estimation
 Breslow Estimate, 468
 Covariate-Adjusted Survival Estimate, 469
 Kalbfleisch-Prentice Estimate, 469
Time-Dependent Covariates, 461
 Wald Tests, 462
Proportional Mean Models, 511
Proportional Odds Alternative, 451
Proportional Odds Model, 357, 523

- Proportional Rate Models, 511
 Quasilikelihood, 585
 GLM Family of Models, 586
 Minimum Chi-Square Estimation, 586
 Overdispersed Poisson Regression Model, 402
- R
- CMPRISK (Competing Risks), 494
 INTERVAL (Interval Censoring), 495
- Radhakrishna Family of Tests, 120, 158, 184, 194
 Random Coefficient Logistic Models, 369
 Random Effects and Mixed Models, 359
 Random Intercept Model, 359
 Random Treatment Effect, 361
- Random Effects Model, 175
 Measurement Error Model, 175
 Poisson Model, 389
 Stratified 2×2 Tables, 175
 Stratified Pair-Matched Tables, 232
 Variance Component Estimate, 179
 Poisson Model, 390
- Rank Tests, 64, 73, 84
 Recombination Fraction, 281
 Recurrent Events, 500
 See also Counting Process and Multiplicative Intensity Model
- Reduced Model Bias, 371
 Relative Risk, 3, 24
 Estimated from Conditional (Retrospective)
 Odds Ratio, 222
 Estimated from Odds Ratio, 204
 Log Relative Risk
 Asymptotic Distribution, 25
 Large Sample Variance, 25, 78
 Matched Pairs, 218, 230
 Matched Retrospective Studies, 222
 Poisson Intensities, 384
 Random Effects Model, 392
 Retrospective, 204
 Residual Variation, 572
 Restricted Alternative Hypothesis, 156
 Test of Association, 156
 Test of Stochastic Ordering, 172
 Ridits, 64
 Right Censoring, 430
 Risk Difference, 23
 Asymptotic Distribution, 23, 78
 Distribution Under the Alternative Hypothesis, 23
 Distribution Under the Null Hypothesis, 23
 Risk Factor, 6, 11
 Robust Inference, 574
 Conditional Logistic Regression Model, 345
 Confidence Limits and Tests, 579
 Correct Model Specification, 575
 Incorrect Model Specification, 576
 Information Sandwich Variance Estimate, 574
- Logistic Regression Model, 317
 Poisson Regression Model, 404
 Proportional Hazards Models, 463
 Score Test, 579
 Wald Test, 579
- Sample Size, 85
 See also Power and Sample Size
 Binomial Distribution with Zero Events, 77
 For Precision of Estimate, 86
 Power and Efficiency, 85
- SAS
- Function CINV, 99
 Function CNONCT, 99
 Function PROBCHI, 99
 ICE Macro (Interval Censoring
 Turnbull Estimate), 495
 PROC CATMOD, 152–153, 235, 288
 PROC FREQ, 34, 48, 135, 188, 233, 235
 AGREE Option, 236
 Matched Pairs, 213
 Rank Tests, 67
 PROC GENMOD, 288, 314, 363, 588
 Least Squares Means (LSMEANS), 314
 Negative Binomial Model, 413
 Negative Binomial Regression Model, 415
 Overdispersed Poisson Model, 402
 Poisson Regression Model, 395–396
 REPEATED Statement, 406
 Type III Option, 315
 Type III Tests, 310
 Zeros-Inflated Poisson Model, 407
- PROC GLIMMIX, 361, 369
 PROC IML, 152
 PROC LIFEREG, 518
 PROC LIFETEST
 Survival Estimation, 446
 Tests of Significance, 455
 PROC LOGISTIC, 287–288
 Class Effects, 300
 TEST Option, 327
 PROC MULTTEST, 71–72
 PROC NLIN, 412
 PROC PHREG, 464, 473
 Stratified Recurrence Models, 512
 PROC Power, 103, 106
 PROC POWER, 486
 PROC Power
 Logistic Regression, 325
 Test for Two Proportions, 106
 Wilcoxon-Mann-Whitney Test, 107
- Scientific Method, 1
 Score Test, 565
 Composite Test, 565
 Conditional Hypergeometric Likelihood, 257
 Conditional Logit Model (Matched Pairs), 268
 Generalized Estimating Equations (GEE), 591

- Logistic Regression Model, 310
- Logit Model, 256
- Mantel-Haenszel Test, 258
- Mantel-Logrank Test in the PH Model, 528
- McNemar's Test, 268
- Poisson Regression Model, 425
- Proportional Hazards Models, 462
- Relative Efficiency versus Likelihood Ratio Test, 567
- Robust, 579
- Stratified-Adjusted Mantel-Logrank Test in the PH Model, 528
- Test of a Subhypothesis: $C(\alpha)$ Tests, 565
- Score-Based Estimate, 258
 - Hazard Ratio, 453
 - Log Odds Ratio, 259
 - Stratified-Adjusted Log Odds Ratio, 261, 279
 - Survival Odds Ratio, 454
- Simpson's Paradox, 137
- Slutsky's Theorem, 543
 - Convergence in Distribution, 543
 - Convergence in Distribution of Transformations, 544
 - Convergence in Probability, 544
- Squared Error Loss, 547, 570
 - Logistic Regression Model, 333
 - Poisson Regression Model, 401
- StatXact, 16, 18, 31, 33–36, 51, 216
- Stochastic Ordering, 172
- Stratification Adjustment and Regression Adjustment, 138
- Stratified Analysis of $R \times C$ Tables
 - Cochran-Armitage Test of Trend, 192
 - Cochran-Mantel-Haenszel Tests, 188
 - Mantel-Haenszel Estimates, 189
 - Mean Scores Test, 191
 - Vector Test of Homogeneity, 191
- Stratified Analysis of 2×2 Tables
 - $C(\alpha)$ Test, 261
 - ARE Versus Unstratified, 112
 - Breslow-Day Test of Homogeneity, 155
 - Cochran's Test of Association, 125, 158, 261
 - Cochran's Test of Homogeneity, 153
 - Conditional Hypergeometric Score Test, 260, 278
 - Contrast Test of Homogeneity, 151
 - DerSimonian and Laird Random Effects Model, 177
 - Logistic Regression Model, 291
 - Mantel-Haenszel Estimates, 126
 - Mantel-Haenszel Test, 124, 260
 - Maximum Likelihood Estimate, 130
 - MVLE, 131, 193
 - Omnibus Test, 148
 - Radbakrisbna Family of Tests, 158
 - Score Test, 260
- Score-Based Estimate of Log Odds Ratio, 261
- Tarone's Test of Homogeneity, 156
- Two Independent Groups, 119
- Zelen's Test of Homogeneity, 154
- Stratified Analysis of Pair-Matched Tables, 227
 - Cochran's Test of Association, 229
 - Cochran's Test of Homogeneity, 229
 - Mantel-Haenszel Analysis, 228
 - Member Stratification, 228
 - MVLE, 229
 - Pair Stratification, 227
- Stratified Analysis of Poisson Intensities, 393
 - Cochran's Test of Homogeneity, 422
 - Efficient Test of Relative Risk, 422
 - MVLE, 422
- Stratified Conditional Hypergeometric Likelihood Maximum Likelihood Estimation, 274, 278
- Stratified Product Binomial Likelihood Maximum Likelihood Estimation of Odds Ratio, 279
- Maximum Likelihood Estimation of Relative Risk, 280
- Maximum Likelihood Estimation of Odds Ratio, 282
- Stratified Recurrence Models, 512
- Subdistribution Function
 - Competing Risk, 493–494
- Suppressor Variable, 138
- Survival Analysis, 430
 - Accelerated Failure-time Model, 497, 516
 - Log-Logistic Model, 518
 - Weibull Model, 517
 - Additional Models, 491
 - Competing Risks, 519
 - Competing Risks Models, 492
 - Counting Process, 500
 - Exponential Regression Models, 496
 - Interval Censoring Models, 495
 - Lehman Alternative, 451
 - Likelihood Function, 431
 - Link-based Models, 495
 - Multiple Event Times, 497
 - Omnibus (MANOVA) Test, 498
 - Robust Covariance Matrix, 498
 - Multiplicative Intensity Model, 507
 - Parametric Models, 496
 - Likelihood Function, 497
 - Proportional Hazards, 451, 519
 - Proportional Mean Models, 511
 - Proportional Odds, 451, 519
 - Proportional Odds Alternative, 451
 - Proportional Odds Model, 523
 - Proportional Rate Models, 511
 - Recurrent Events, 499
 - Stratified Recurrence Models, 512
 - Weibull Regression Models, 496

- Survival Distribution, 430
 Exponential, 430, 483, 514
 Weibull, 514
 Survival Function
 Actuarial Estimate, 444
 Binomial Variance Versus Large Sample Variance, 438, 522
 Comparison of Two Groups, 436
 Discrete or Grouped Time, 444, 521
 Kaplan-Meier Estimate, 432
 Large Sample Variance of Log Survival, 434, 436
 Large Sample Variance of Logit of Survival, 522
 Large Sample Variance of Log-Log Survival, 435–436, 522
 Modified Kaplan-Meier Estimate, 444
 Nelson-Aalen Estimate, 436
 Proportional Hazards Models
 Breslow Estimate, 468
 Covariate-Adjusted Survival Estimate, 469
 Kalbfleisch-Prentice Estimate, 469
 Synergism, 138
 Tarone's Test of Homogeneity of Odds Ratios, 156
 Tarone-Ware Family of Tests for Survival Functions, 452
 Taylor's Approximation, 541
 Test of Homogeneity, 150
 Contrast Test, 151
 Null and Alternative Hypotheses, 150
 Test of Partial Association
 Null and Alternative Hypotheses, 150, 156–157
 Test-Based Confidence Limits, 18, 77
 Binomial Distribution Probability, 18
 Mantel-Haenszel Stratified-Adjusted Odds Ratio, 127
 Odds Ratio, 42
 Time-Dependent Covariate, 11
 Two-Stage Model, 177
 Measurement Error Model, 177
 Poisson Model, 389
 Type I and II Errors, 87
 Type I Error Probability, 37, 44, 87–88
 Type II Error Probability, 87–88
 Type III Tests in SAS PROC GENMOD, 310, 315
 Uncertainty Coefficient, 49, 376
 University Group Diabetes Program/524 Variance Component, 86
 Violations of the Homogeneous Poisson Assumptions, 388
 Wald Test, 561
 T^2 -Like Test, 62, 83, 146
 Caveats, 313
 Composite Test, 562
 Contrast Test of Homogeneity, 152
 Element-wise Tests, 562
 Logistic Regression Model, 312
 Logit Model, 255
 Proportional Hazards Models, 462
 Robust, 579
 Test of A Linear Hypothesis, 562
 Weibull Survival Distribution, 514
 Accelerated Failure-time Model, 517
 Maximum Likelihood Estimation, 514
 Weighted Least Squares (WLS), 548
 Weighted Mantel-Haenszel Test, 449
 See also Aalen-Gill Test Statistics
 G^p Family, 452
 Lehman Alternative, 451
 Logrank Test, 450
 Measures of Association, 453
 Optimal Weights, 450
 Proportional Odds Alternative, 451
 Score-Based Estimate of Hazard Ratio, 453
 Score-Based Estimate of Survival Odds Ratio, 454
 Stratified-Adjusted, 453
 Tarone-Ware Family, 452
 Weights Inversely Proportional to the Variances, 114, 118, 131, 134, 158
 Wei-Lachin Test of Stochastic Ordering, 171
 Chi-Square Test, 173
 Null and Alternative Hypotheses, 172
 Stratified Pair Matching, 231
 Wilcoxon Rank Sum Test, 64
 Wilcoxon Test for Lifetables
 Aalen-Gill Family Test, 506
 Gehan Test, 451
 Peto-Peto-Prentice Test, 451, 524
 Woolf's Variance Estimate, 16, 27, 31, 78
 Working Correlation Matrix, 589
 Exchangeable, 590
 First-order Autoregressive, 590
 Independence, 590
 Unstructured, 590
 Zelen's Test of Homogeneity, 154
 Zeros-Inflated Poisson (ZIP) Regression Model, 407
 Z-Test
 A Proportion, 15
 Functions of Two Proportions, 80
 General, 37
 Matched Pairs (McNemar's Test), 213
 Null Versus Alternative Variance, 38
 Poisson Intensities of Two Populations, 385
 Poisson Intensities, Random Effects Model, 392
 Survival Probabilities for Two Groups, 437
 Two Independent Proportions, 38, 116
 Two Means, 116
 Two Poisson Intensities, 117

Biostatistical Methods: The Assessment of Relative Risks, Second Edition
by John M. Lachin
Copyright © 2011 John Wiley & Sons, Inc.

WILEY SERIES IN PROBABILITY AND STATISTICS
ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

The **Wiley Series in Probability and Statistics** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

- † ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Second Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
- † BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
- BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
- BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*
- BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
- BERNARDO and SMITH · Bayesian Theory
- BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
- BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
- BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
- BILLINGSLEY · Probability and Measure, *Third Edition*
- BIRKES and DODGE · Alternative Methods of Regression
- BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics
- BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
- BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
- BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
- BOLLEN · Structural Equations with Latent Variables
- BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
- BOROVKOV · Ergodicity and Stability of Stochastic Processes
- BOULEAU · Numerical Methods for Stochastic Processes
- BOX · Bayesian Inference in Statistical Analysis
- BOX · R. A. Fisher, the Life of a Scientist
- BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*
- * BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
- BOX and FRIENDS · Improving Almost Anything, *Revised Edition*
- BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Edition*
- BOX, JENKINS, and REINSEL · Time Series Analysis: Forecasting and Control, *Fourth Edition*
- BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, *Second Edition*
- † BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
- BROWN and HOLLANDER · Statistics: A Biomedical Introduction
- BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
- BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
- CAIROLI and DALANG · Sequential Stochastic Optimization
- CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
- CHAN · Time Series: Applications to Finance with R and S-Plus®, *Second Edition*
- CHARALAMBIDES · Combinatorial Methods in Discrete Distributions
- CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*
- CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
- CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
- CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
- CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
- CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- CLARKE · Linear Models: The Theory and Application of Analysis of Variance
 CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
- * COCHRAN and COX · Experimental Designs, *Second Edition*
- COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences
- CONGDON · Applied Bayesian Modelling
- CONGDON · Bayesian Models for Categorical Data
- CONGDON · Bayesian Statistical Modelling
- CONOVER · Practical Nonparametric Statistics, *Third Edition*
- COOK · Regression Graphics
- COOK and WEISBERG · Applied Regression Including Computing and Graphics
- COOK and WEISBERG · An Introduction to Regression Graphics
- CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
- COVER and THOMAS · Elements of Information Theory
- COX · A Handbook of Introductory Statistical Methods
- * COX · Planning of Experiments
- CRESSIE · Statistics for Spatial Data, *Revised Edition*
- CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
- DANIEL · Applications of Statistics to Industrial Experimentation
- DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
- * DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
- DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
- DAVID and NAGARAJA · Order Statistics, *Third Edition*
- * DEGROOT, FIENBERG, and KADANE · Statistics and the Law
- DEL CASTILLO · Statistical Process Adjustment for Quality Control
- DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
- DEMIDENKO · Mixed Models: Theory and Applications
- DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
- DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
- DEY and MUKERJEE · Fractional Factorial Plans
- DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
- DODGE · Alternative Methods of Regression
- * DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
- * DOOB · Stochastic Processes
- DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
- DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
- DRYDEN and MARDIA · Statistical Shape Analysis
- DUDEWICZ and MISHRA · Modern Mathematical Statistics
- DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
- DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
- EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
- * ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
- ENDERS · Applied Econometric Time Series
- † ETHIER and KURTZ · Markov Processes: Characterization and Convergence
- EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- FELLER · An Introduction to Probability Theory and Its Applications, Volume I,
Third Edition, Revised; Volume II, *Second Edition*
- FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
- FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis
- * FLEISS · The Design and Analysis of Clinical Experiments
- FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
- † FLEMING and HARRINGTON · Counting Processes and Survival Analysis
- FUKIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations
- FULLER · Introduction to Statistical Time Series, *Second Edition*
- † FULLER · Measurement Error Models
- GALLANT · Nonlinear Statistical Models
- GEISSER · Modes of Parametric Statistical Inference
- GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
- GEWEKE · Contemporary Bayesian Econometrics and Statistics
- GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
- RIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
- GIFI · Nonlinear Multivariate Analysis
- GIVENS and HOETING · Computational Statistics
- GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
- GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
- GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
- GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
- GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*
- GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*
- * HAHN and SHAPIRO · Statistical Models in Engineering
- HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
- HALD · A History of Probability and Statistics and their Applications Before 1750
- HALD · A History of Mathematical Statistics from 1750 to 1930
- † HAMEL · Robust Statistics: The Approach Based on Influence Functions
- HANNAN and DEISTLER · The Statistical Theory of Linear Systems
- HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications
- HEIBERGER · Computation for the Analysis of Designed Experiments
- HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
- HEDEKER and GIBBONS · Longitudinal Data Analysis
- HELLER · MACSYMA for Statisticians
- HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*
- HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design
- HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance
- * HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
- * HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
- HOCHBERG and TAMHANE · Multiple Comparison Procedures
- HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
 HOGG and KLUGMAN · Loss Distributions
 HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
 HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
 HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling
 of Time-to-Event Data, *Second Edition*
[†] HUBER and RONCHETTI · Robust Statistics, *Second Edition*
 HUBERTY · Applied Discriminant Analysis
 HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis,
 Second Edition
 HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*
 HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory
 and Practice
 HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—
 with Commentary
 HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
 IMAN and CONOVER · A Modern Approach to Statistics
[†] JACKSON · A User's Guide to Principle Components
 JOHN · Statistical Methods in Engineering and Quality Assurance
 JOHNSON · Multivariate Statistical Simulation
 JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A
 Volume in Honor of Samuel Kotz
 JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, *Fifth Edition*
 JOHNSON and KOTZ · Distributions in Statistics
 JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the
 Seventeenth Century to the Present
 JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
 Volume 1, *Second Edition*
 JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
 Volume 2, *Second Edition*
 JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
 JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*
 JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of
 Econometrics, *Second Edition*
 JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations
 JUREK and MASON · Operator-Limit Distributions in Probability Theory
 KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
 KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
 KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second
 Edition*
 KARIYA and KURATA · Generalized Least Squares
 KASS and VOS · Geometrical Foundations of Asymptotic Inference
[†] KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster
 Analysis
 KEDEM and FOKIANOS · Regression Models for Time Series Analysis
 KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
 KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
 KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
 KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
 KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization
 KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions,
 Third Edition
 KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models:
 From Data to Decisions, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
- KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
- KOWALSKI and TU · Modern Applied U-Statistics
- KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation
- KROONENBERG · Applied Multiway Data Analysis
- KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering
- LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*
- LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
- LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
- LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
- LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
- LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
- LAWSON · Statistical Methods in Spatial Epidemiology
- LE · Applied Categorical Data Analysis
- LE · Applied Survival Analysis
- LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
- LEPAGE and BILLARD · Exploring the Limits of Bootstrap
- LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
- LIAO · Statistical Group Comparison
- LINDVALL · Lectures on the Coupling Method
- LIN · Introductory Stochastic Analysis for Finance and Insurance
- LINHART and ZUCCHINI · Model Selection
- LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
- LLOYD · The Statistical Analysis of Categorical Data
- LOWEN and TEICH · Fractal-Based Point Processes
- MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
- MALLER and ZHOU · Survival Analysis with Long Term Survivors
- MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
- MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
- MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
- MARCHETTE · Random Graphs for Statistical Pattern Recognition
- MARDIA and JUPP · Directional Statistics
- MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
- McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*
- McFADDEN · Management of Data in Clinical Trials, *Second Edition*
- * McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
- McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
- McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*
- McLACHLAN and PEEL · Finite Mixture Models
- McNEIL · Epidemiological Research Methods
- MEEKER and ESCOBAR · Statistical Methods for Reliability Data
- MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*
- * MILLER · Survival Analysis, *Second Edition*
- MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting
- MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fourth Edition*
- MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
- MUIRHEAD · Aspects of Multivariate Statistical Theory
- MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
- MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
- MURTHY, XIE, and JIANG · Weibull Models
- MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*
- MYERS, MONTGOMERY, VINING, and ROBINSON · Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*
- † NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
- † NELSON · Applied Life Data Analysis
- NEWMAN · Biostatistical Methods in Epidemiology
- OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
- OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
- OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
- PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
- PANJER · Operational Risk: Modeling and Analytics
- PANKRATZ · Forecasting with Dynamic Regression Models
- PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
- * PARZEN · Modern Probability Theory and Its Applications
- PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
- PIANTADOSI · Clinical Trials: A Methodologic Perspective
- PORT · Theoretical Probability for Applications
- POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
- POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality
- PRESS · Bayesian Statistics: Principles, Models, and Applications
- PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
- PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
- PUKELSHEIM · Optimal Experimental Design
- PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
- † PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
- QIU · Image Processing and Jump Regression Analysis
- * RAO · Linear Statistical Inference and Its Applications, *Second Edition*
- RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
- RENCHER · Linear Models in Statistics
- RENCHER · Methods of Multivariate Analysis, *Second Edition*
- RENCHER · Multivariate Statistical Inference with Applications
- * RIPLEY · Spatial Statistics
- * RIPLEY · Stochastic Simulation
- ROBINSON · Practical Strategies for Experimenting

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
 ROLSKI, SCHMIDL, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
- ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
 ROSS · Introduction to Probability and Statistics for Engineers and Scientists
 ROSSI, ALLENBY, and MCCULLOCH · Bayesian Statistics and Marketing
- † ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
 * RUBIN · Multiple Imputation for Nonresponse in Surveys
 RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*
 RUBINSTEIN and MELAMED · Modern Simulation and Modeling
 RYAN · Modern Engineering Statistics
 RYAN · Modern Experimental Design
 RYAN · Modern Regression Methods, *Second Edition*
 RYAN · Statistical Methods for Quality Improvement, *Second Edition*
 SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
- * SCHEFFE · The Analysis of Variance
 SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
 SCHOTT · Matrix Analysis for Statistics, *Second Edition*
 SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
 SCHUSS · Theory and Applications of Stochastic Differential Equations
 SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
- † SEARLE · Linear Models for Unbalanced Data
 † SEARLE · Matrix Algebra Useful for Statistics
 † SEARLE, CASELLA, and MCCULLOCH · Variance Components
 SEARLE and WILLETT · Matrix Algebra for Applied Economics
 SEBER · A Matrix Handbook For Statisticians
 † SEBER · Multivariate Observations
 SEBER and LEE · Linear Regression Analysis, *Second Edition*
 † SEBER and WILD · Nonlinear Regression
 SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
- * SERFLING · Approximation Theorems of Mathematical Statistics
 SHAFER and VOVK · Probability and Finance: It's Only a Game!
 SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
 SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
 SRIVASTAVA · Methods of Multivariate Statistics
 STAPLETON · Linear Statistical Models, *Second Edition*
 STAPLETON · Models for Probability and Statistical Inference: Theory and Applications
 STAUDTE and SHEATHER · Robust Estimation and Testing
 STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
 STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
 STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods
 STYAN · The Collected Papers of T. W. Anderson: 1943–1985
 SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
 TAKEZAWA · Introduction to Nonparametric Regression
 TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications
 TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
 THOMPSON · Empirical Model Building
 THOMPSON · Sampling, *Second Edition*
 THOMPSON · Simulation: A Modeler's Approach

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- THOMPSON and SEBER · Adaptive Sampling
- THOMPSON, WILLIAMS, and FINLAY · Models for Investors in Real World Markets
- TCIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
- TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
- TSAY · Analysis of Financial Time Series, *Third Edition*
- UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
- † VAN BELLE · Statistical Rules of Thumb, *Second Edition*
- VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
- VESTRUP · The Theory of Measures and Integration
- VIDAKOVIC · Statistical Modeling by Wavelets
- VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
- WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
- WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
- WEISBERG · Applied Linear Regression, *Third Edition*
- WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons
- WELSH · Aspects of Statistical Inference
- WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
- WHITTAKER · Graphical Models in Applied Multivariate Statistics
- WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
- WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
- WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
- WOODWORTH · Biostatistics: A Bayesian Introduction
- WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
- WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*
- WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis
- YANG · The Construction Theory of Denumerable Markov Processes
- YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
- ZACKS · Stage-Wise Adaptive Designs
- ZELTERMAN · Discrete Distributions—Applications in the Health Sciences
- * ZELLNER · An Introduction to Bayesian Inference in Econometrics
- ZHOU, MCCLISH, and OBUCHOWSKI · Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.