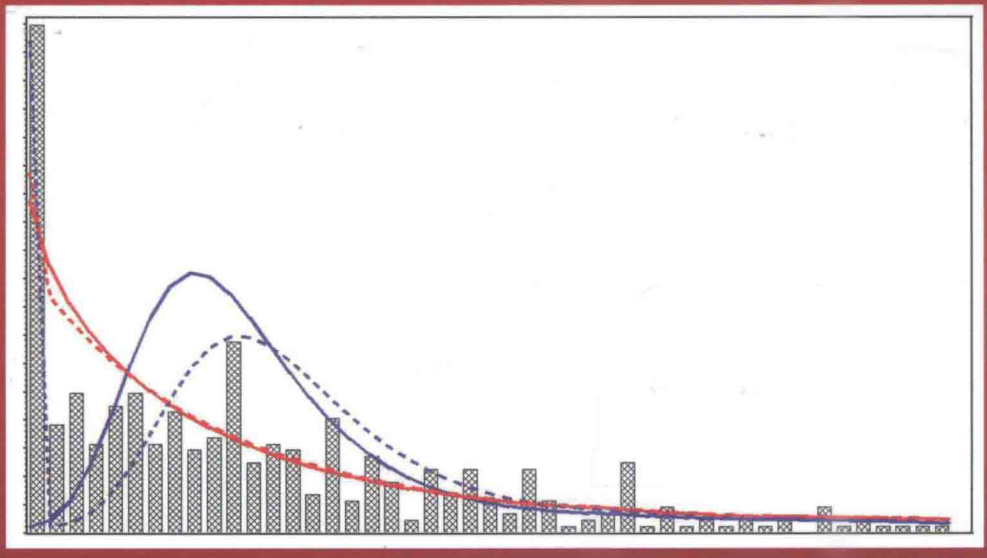
Texts in Statistical Science

Applied Categorical and Count Data Analysis



Wan Tang
Hua He
Xin M. Tu

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Texts in Statistical Science

Applied Categorical and Count Data Analysis

Wan Tang
Hua He
Xin M. Tu


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
Version Date: 20120412

International Standard Book Number: 978-1-4398-0624-1 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
P. J. Smith

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

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

Applied Categorical and Count Data Analysis
W. Tang, H. He, and X.M. Tu

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

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

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

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Bayesian Data Analysis, Second Edition
A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

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

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Design and Analysis of Experiments with SAS
J. Lawson

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

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis, Second Edition
M. Woodward

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

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Extending the Linear Model with R — Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Generalized Additive Models: An Introduction with R
S. Wood

Graphics for Statistics and Data Analysis with R
K.J. Keen

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

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

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

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

Introduction to Probability with R
K. Baclawski

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Introduction to Statistical Limit Theory
A.M. Polansky

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

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

Linear Models with R
J.J. Faraway

Logistic Regression Models
J.M. Hilbe

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

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

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

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

Multivariate Survival Analysis and Competing Risks
M. Crowder

Pólya Urn Models
H. Mahmoud

Practical Data Analysis for Designed Experiments
B.S. Yandell

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

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Principles of Uncertainty
J.B. Kadane

Probability — Methods and Measurement
A. O'Hagan

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

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

Readings in Decision Analysis
S. French

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

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

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

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

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

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

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

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

Statistics in Research and Development, Second Edition
R. Caulcutt

Stochastic Processes: An Introduction, Second Edition
P.W. Jones and P. Smith

Survival Analysis Using S — Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Applied Categorical and Count Data Analysis

Preface

This book focuses on statistical analysis of discrete data, including categorical and count outcomes. Discrete variables are abundant in practice, and knowledge about and ability to analyze such data is important for professionals and practitioners in a wide range of biomedical and psychosocial research areas. Although there are some excellent books on this general subject such as those by Agresti (2002, 2007), Long (1997), Long and Freese (2006), and Stokes et al. (2009), a book that includes models for longitudinal data, real data examples with detailed programming codes, as well as intuitive explanations of the models and their interpretations and differences thereupon will complement the repertoire of existing texts. Motivated by the lack of such a text, we decided to write this book five years ago when preparing a graduate-level biostatistics course on this topic for students within a medical school setting at the University of Rochester. The lecture notes from which this book has evolved have been used for the course over the past five years.

In addition to the classic concepts such as contingency tables and popular topics such as logistic and Poisson regression models, as covered by most available textbooks on categorical data analysis, this book also includes many modern topics. These include models for zero modified count outcomes, longitudinal data analysis (both parametric and semiparametric), reliability analysis, and popular methods for dealing with missing values. More importantly, programming codes are provided for all the examples in the book for the four major software packages, R, SAS, SPSS, and Stata, so that when reading the examples readers can immediately put their knowledge into practice by trying out the codes with the data in the examples using the statistical packages of their choice, and/or adapt and even extend them to fit settings arising from their own studies.

We view effective learning as a process of "reverse engineering" in the sense that one develops an in-depth appreciation of a concept, model, or approach by tracing its bumble beginnings that motivate its development in the first place. With this philosophy in mind, we try to describe the basic ideas underlying each concept, model, and approach introduced in this book so that even without rigorous mathematical arguments, readers can have a good grasp of the fundamentals of the concept and methodology. For the rather technical-savvy audience, we have also included a section in Chapter 1 to review some key results on statistical inference to help facilitate the discussion and understanding of the theoretical aspects of the models and inference methods introduced in the subsequent chapters, complemented by theory-oriented ex-

ercises at the end of each chapter. Readers should not be discouraged by such theoretical materials and exercises, since skipping such theoretical justifications will not hamper understanding of the concepts and models and principles of applying them in practice. The book is pretty much self-contained, with no prerequisite for using this book, although knowledge on statistics in general is helpful. Fundamental concepts such as confidence intervals, hypothesis tests, and p-values are briefly introduced as they first appear in the text so that people without former exposure to statistics may still benefit from the book.

The outline of the book is as follows. In addition to the review section mentioned above, Chapter 1 also presents various types of discrete random variables, together with an introduction of the study data that will be used throughout the book.

In Chapter 2, we first study individual random variables and introduce the popular discrete distributions including the binomial, multinomial, and Poisson models. Next we concentrate on the study of relationship between two categorical variables, i.e., the study of two-way contingency tables. This is followed in Chapter 3 by stratified two-way tables, controlling for potential categorical confounding variables.

When there are more than two categorical variables, or there are continuous variables present, regression analysis becomes necessary to study the relationship between such variables. In Chapter 4, we introduce regression models for categorical responses. We first discuss logistic regression for binary responses in detail, including methods to reduce bias for relatively small samples such as exact logistic models. Less popular models for binary responses such as the Probit and complementary log-log models are then discussed, followed by the models for general polytomous categorical outcomes to conclude this chapter.

Chapter 5 focuses on regression analysis of count responses. As the most commonly used models in this setting, the Poisson log-linear regression is first studied in detail, followed by a discussion on overdispersion, a common violation of the Poisson model, along with its detection and correction within the confines of this model using robust inference methods, such as the sandwich variance estimate. Alternative models that explicitly account for the sources of overdispersion and structural zero, another common violation of the Poisson, such as the negative binomial, hurdle, and zero-modified models, are then introduced to formally address such deviations from the Poisson. This chapter concludes with a systematic guide to modeling count responses using the different models introduced. Chapter 6 illustrates a major application of the Poisson log-linear regression, as it applies to general contingency tables to facilitate inference about the relationship between multiple variables, which is algebraically too complex using the classic methods discussed in Chapters 2 and 3. Also included in Chapter 6 is a section on model selection that introduces popular criteria for deriving optimal models within a given context.

Chapter 7 discusses analyses for discrete survival times. Survival analysis is widely used in statistical applications involving time to occurrence of some event of interest such as heart attacks and suicide attempts. We discuss non-

parametric life table methods as well as regression approaches.

The statistical methods covered in Chapters 2-7 are mainly for cross-sectional studies, where the data only include a single assessment point for every subject. This is not the case for longitudinal studies, where the same set of outcomes such as disease status is repeatedly measured from the same subject over time. Methods for longitudinal data must address the within-subject correlations in repeatedly measured outcomes over time. In Chapter 8, we introduce longitudinal data and models for such data, and focus on the popular parametric mixed-effects models and semiparametric generalized estimating equations.

Chapter 9 discusses validity and reliability analysis for diagnostic tests and measuring instruments. We discuss how to assess the accuracy of an ordinal test when the true status is known, using the theory of receiver operating characteristics (ROC) curves. We introduce measurement error models for assessing latent constructs, and discuss popular indices for addressing inter-rater agreement and instrument validity and reliability such as Cronbach's alpha coefficient and Kappa.

In Chapter 10, we discuss how to deal with missing values. Common approaches such as multiple imputation and inverse probability weighting methods are introduced. Since applications of the missing value concept really go beyond addressing the problem of missing values in study outcomes, we also illustrate how to apply the principles of such methods to a range of seemingly unrelated issues such as causal inference and survey sampling.

This book can serve as a primary text for a course on categorical and count data analysis for senior undergraduate, beginning as well as senior graduate students in biostatistics. It also serves well as a self-learning text for biomedical and psychosocial researchers interested in this general subject. Based on our own experiences, Chapters 1 through 7 can be covered in a one-semester course.

We would like to express our appreciation to all who have contributed to this book. We would like thank the students at the University of Rochester who took the course in the past five years, many of whom have provided countless helpful comments and feedbacks. We would also like to thank Dr. Yinglin Xia and Dr. Guoxin Zuo, who proofed many parts of the book, and offered numerous valuable comments and suggestions; Dr. Naiji Lu, who helped with some of the examples in Chapter 9 whose analyses are not supported by standard software packages; and Dr. Jun Hu, who proofread the entire book multiple times to help eradicate errors and typos. We are grateful to Drs. Linda Chaudron, Steve Lamberti, Jeffrey Lyness, Mary Caserta, and Paul Duberstein from the University of Rochester, and Dr. Dianne Morrison-Beedy from the University of South Florida for graciously sharing their study data for use in the book as real data examples. We are also thankful to editor David Grubbs for his patience and continuing support despite multiple delays on the project on our part, to one anonymous reviewer for his/her critical comments and constructive suggestions that have led to an improved presentation, and

to staffs at CRC who carefully proofread the manuscript and helped with some technique issues and numerous corrections. And last but not least, we thank all the faculty and staff in the department for their support.

Contents

List of Tables xiii

List of Figures xv

Preface xvii

1 Introduction 1
- 1.1 Discrete Outcomes . . . 1
- 1.2 Data Source . . . 3
- 1.3 Outline of the Book . . . 4
  - 1.3.1 Distribution of Random Variables . . . 5
  - 1.3.2 Association between Two Random Variables . . . 5
  - 1.3.3 Regression Analysis . . . 7
  - 1.3.4 Log-Linear Methods for Contingency Tables . . . 8
  - 1.3.5 Discrete Survival Data Analysis . . . 9
  - 1.3.6 Longitudinal Data Analysis . . . 10
  - 1.3.7 Validity and Reliability Data Analysis . . . 12
  - 1.3.8 Incomplete Data Analysis . . . 13
- 1.4 Review of Key Statistical Results . . . 14
  - 1.4.1 Central Limit Theorem and Law of Large Numbers . . . 15
  - 1.4.2 Delta Method and Slutsky's Theorem . . . 18
  - 1.4.3 Maximum Likelihood Estimate . . . 19
  - 1.4.4 Estimating Equations . . . 22
  - 1.4.5 U-Statistics . . . 24
- 1.5 Software . . . 26

Exercises . . . 27

2 Contingency Tables 31
- 2.1 Inference for One-Way Frequency Table . . . 31
  - 2.1.1 Binary Case . . . 32
  - 2.1.2 Inference for Multinomial Variable . . . 37
  - 2.1.3 Inference for Count Variable . . . 39
- 2.2 Inference for 2×2 Table . . . 43
  - 2.2.1 Testing Association . . . 45
  - 2.2.2 Measures of Association . . . 50
  - 2.2.3 Test for Marginal Homogeneity . . . 55
  - 2.2.4 Agreement . . . 56

2.3 Inference for $2 \times r$ Tables 59
2.3.1 Cochran–Armitage Trend Test 60
2.3.2 Mann–Whitney–Wilcoxon Test 61
2.4 Inference for $s \times r$ Table 64
2.4.1 Tests of Association 65
2.4.2 Marginal Homogeneity and Symmetry 71
2.4.3 Agreement 73
2.5 Measures of Association 76
2.5.1 Measures of Association for Ordinal Outcome 76
2.5.2 Measures of Association for Nominal Outcome 83
Exercises 87

3 Sets of Contingency Tables **93**
3.1 Confounding Effects 94
3.2 Sets of 2×2 Tables 97
3.2.1 Cochran–Mantel–Haenszel Test for Independence . . . 99
3.2.2 Estimates and Tests of Common Odds Ratios 101
3.3 Sets of $s \times r$ Tables 106
3.3.1 Tests of General Association 106
3.3.2 Mean Score Statistic 109
3.3.3 Correlation Statistic 110
3.3.4 Kappa Coefficients for Stratified Tables 111
Exercises 113

4 Regression Models for Categorical Response **115**
4.1 Logistic Regression for Binary Response 116
4.1.1 Motivation of Logistic Regression 116
4.1.2 Definition of Logistic Models 117
4.1.3 Parameter Interpretation 120
4.1.4 Invariance to Study Designs 123
4.1.5 Simpson's Paradox Revisited 125
4.1.6 Breslow–Day Test and Moderation Analysis 127
4.2 Inference About Model Parameters 130
4.2.1 Maximum Likelihood Estimate 130
4.2.2 General Linear Hypotheses 132
4.2.3 Exact Inference for Logistic Regression 139
4.2.4 Bias Reduced Logistic Regression 145
4.3 Goodness of Fit 147
4.3.1 The Pearson Chi-Square Statistic 148
4.3.2 The Deviance Test 151
4.3.3 The Hosmer–Lemeshow Test 152
4.3.4 Lack of Fit 154
4.4 Generalized Linear Models 155
4.4.1 Introduction 155
4.4.2 Regression Models for Binary Response 156

4.4.3 Inference . . . 158
4.5 Regression Models for Polytomous Response . . . 159
4.5.1 Model for Nominal Response . . . 160
4.5.2 Models for Ordinal Response . . . 162
4.5.3 Inference . . . 165
Exercises . . . 168

5 Regression Models for Count Response 173
5.1 Poisson Regression Model for Count Response . . . 173
5.1.1 Parameter Interpretation . . . 174
5.1.2 Inference About Model Parameters . . . 175
5.1.3 Offsets in Log-Linear Model . . . 177
5.2 Goodness of Fit . . . 178
5.2.1 Pearson's Chi-Square Statistic . . . 179
5.2.2 Scaled Deviance Statistic . . . 180
5.3 Overdispersion . . . 182
5.3.1 Detection of Overdispersion . . . 182
5.3.2 Correction for Overdispersion . . . 183
5.4 Parametric Models for Clustered Count Response . . . 187
5.4.1 Negative Binomial Model . . . 187
5.4.2 Zero-Modified Poisson and Negative Binomial Models 190
5.4.3 Zero-Truncated Poisson and NB Regression Models . . 196
5.4.4 Hurdle Models . . . 197
Exercises . . . 199

6 Log-Linear Models for Contingency Tables 201
6.1 Analysis of Log-Linear Models . . . 202
6.1.1 Motivation . . . 202
6.1.2 Log-Linear Models for Contingency Tables . . . 204
6.1.3 Parameter Interpretation . . . 204
6.1.4 Inference . . . 206
6.2 Two-Way Contingency Tables . . . 208
6.2.1 Independence . . . 208
6.2.2 Symmetry and Marginal Homogeneity . . . 211
6.3 Three-Way Contingency Tables . . . 212
6.3.1 Independence . . . 213
6.3.2 Association Homogeneity . . . 216
6.4 Irregular Tables . . . 218
6.4.1 Structure Zeros in Contingency Tables . . . 219
6.4.2 Models for Irregular Tables . . . 221
6.4.3 Bradley–Terry Model . . . 223
6.5 Model Selection . . . 225
6.5.1 Model Evaluation . . . 225
6.5.2 Stepwise Selection . . . 226
6.5.3 Graphical Models . . . 231

Exercises . . . 232

7 Analyses of Discrete Survival Time **237**
7.1 Special Features of Survival Data . . . 237
7.1.1 Censoring . . . 238
7.1.2 Truncation . . . 239
7.1.3 Discrete Survival Time . . . 240
7.1.4 Survival and Hazard Functions . . . 242
7.2 Life Table Methods . . . 243
7.2.1 Life Tables . . . 244
7.2.2 The Mantel–Cox Test . . . 248
7.3 Regression Models . . . 250
7.3.1 Complementary Log-Log Regression . . . 250
7.3.2 Discrete Proportional Odds Model . . . 253
Exercises . . . 254

8 Longitudinal Data Analysis **257**
8.1 Data Preparation and Exploration . . . 259
8.1.1 Longitudinal Data Formats . . . 259
8.1.2 Exploratory Analysis . . . 261
8.2 Marginal Models . . . 264
8.2.1 Models for Longitudinal Data . . . 265
8.2.2 Generalized Estimation Equations . . . 266
8.2.3 Extensions to Categorical Responses . . . 274
8.3 Generalized Linear Mixed-Effects Model . . . 275
8.3.1 Linear Mixed-Effects Models . . . 276
8.3.2 Generalized Linear Mixed-Effects Models . . . 278
8.3.3 Comparison of GLMM with Marginal Models . . . 280
8.3.4 Maximum Likelihood Inference . . . 281
8.4 Model Diagnostics . . . 282
8.4.1 Marginal Models . . . 282
8.4.2 Generalized Linear Mixed-Effect Models . . . 284
Exercises . . . 285

9 Evaluation of Instruments **289**
9.1 Diagnostic-Ability . . . 290
9.1.1 Receiver Operating Characteristic Curves . . . 290
9.1.2 Inference . . . 294
9.1.3 Areas under ROC Curves . . . 295
9.2 Criterion Validity . . . 297
9.2.1 Concordance Correlation Coefficient . . . 298
9.3 Internal Reliability . . . 299
9.3.1 Spearman–Brown Rho . . . 301
9.3.2 Cronbach Coefficient Alpha . . . 302
9.3.3 Intraclass Correlation Coefficient . . . 304

9.4 Test-Retest Reliability 306
Exercises 308

10 Analysis of Incomplete Data **311**
10.1 Incomplete Data and Associated Impact 311
10.1.1 Observational Missing 311
10.1.2 Missing by Design 312
10.1.3 Counterfactual Missing 313
10.1.4 Impact of Missing Values 313
10.2 Missing Data Mechanism 315
10.2.1 Missing Completely at Random 315
10.2.2 Missing at Random 316
10.2.3 Missing Not at Random 319
10.3 Methods for Incomplete Data 320
10.3.1 Maximum Likelihood Method 320
10.3.2 Imputation Methods 322
10.3.3 Inverse Probability Weighting 327
10.3.4 Sensitivity Analysis 328
10.4 Applications 329
10.4.1 Verification Bias of Diagnostic Studies 330
10.4.2 Causal Inference of Treatment Effects 333
10.4.3 Longitudinal Data with Missing Values 336
10.4.4 Survey Studies 341
Exercises 344

References **347**

Index **359**

List of Tables

1.1 Gender by MS for the Metabolic Syndrome study 6
1.2 First major depression diagnosis (dropout) post baseline . . . 10
1.3 First major depression diagnosis for men (women) 10

2.1 Depression diagnosis in the DOS study 38
2.2 Frequency of protected vaginal sex 42
2.3 A typical 2×2 contingency table 45
2.4 Recidivism before and after treatment 47
2.5 Table probabilities 50
2.6 Depression of patients at years 0 and 1 (DOS study) 56
2.7 Depression diagnoses based on the probands and informants . 58
2.8 Depression diagnoses by gender in the DOS study 61
2.9 Depression diagnoses based on the probands and informants . 73

3.1 Success rates of two hospitals 94
3.2 Success rates of two hospitals stratified by disease severity . . 95
3.3 A set of q 2×2 tables 97
3.4 Depression by gender, stratified by education 101
3.5 Depression diagnosis, stratified by informant gender 113

4.1 A 2×2 contingency table for a prospective study 123

6.1 Distribution of pre- and postweight categories 221

8.1 Horizontal format for longitudinal data 260
8.2 Vertical format for longitudinal data 261
8.3 Proportions of major depression at each visit 262

9.1 PM correlation and CCC between recall (y_{1i}) and diary (y_{2i}) 299
9.2 Cronbach coefficient alpha and ICC for the PF domain 306
9.3 PM correlation, CCC, and ICC between admission and 1-2 days post admission to hospital for the PF domain 307

10.1 A hypothetical study of a diagnostic test 314
10.2 Estimates of prevalences under different γ's 329
10.3 GEE, WGEE, and MI-GEE estimates of β_3 340

List of Figures

1.1 HIV infection to AIDS onset. 11

3.1 Mean responses of a continuous outcome for two hospitals. . . 96

5.1 Distribution of VCD. 190

8.1 HIV knowledge scores of a random sample. 258
8.2 Proportions of major depression and confidence intervals. . . 263

9.1 Binormal ROC curves. 292
9.2 Empirical ROC curves for EPDS and BDI II. 297