

Contents

Preface	xiii
I Getting Started	1
1 Introduction	3
1.1 Data visualization and categorical data: Overview	3
1.2 What is categorical data?	4
1.2.1 Case form vs. frequency form	5
1.2.2 Frequency data vs. count data	6
1.2.3 Univariate, bivariate, and multivariate data	7
1.2.4 Explanatory vs. response variables	7
1.3 Strategies for categorical data analysis	8
1.3.1 Hypothesis testing approaches	8
1.3.2 Model building approaches	10
1.4 Graphical methods for categorical data	13
1.4.1 Goals and design principles for visual data display	13
1.4.2 Categorical data require different graphical methods	17
1.4.3 Effect ordering and rendering for data display	18
1.4.4 Interactive and dynamic graphics	21
1.4.5 Visualization = Graphing + Fitting + Graphing	22
1.4.6 Data plots, model plots, and data+model plots	25
1.4.7 The 80–20 rule	26
1.5 Chapter summary	28
1.6 Lab exercises	29
2 Working with Categorical Data	31
2.1 Working with R data: vectors, matrices, arrays, and data frames	32
2.1.1 Vectors	32
2.1.2 Matrices	33
2.1.3 Arrays	35
2.1.4 Data frames	37
2.2 Forms of categorical data: case form, frequency form, and table form	39
2.2.1 Case form	39

2.2.2	Frequency form	40
2.2.3	Table form	41
2.3	Ordered factors and reordered tables	43
2.4	Generating tables: table and xtabs	44
2.4.1	table()	44
2.4.2	xtabs()	46
2.5	Printing tables: structable and ftable	47
2.5.1	Text output	47
2.6	Subsetting data	48
2.6.1	Subsetting tables	48
2.6.2	Subsetting structables	49
2.6.3	Subsetting data frames	50
2.7	Collapsing tables	51
2.7.1	Collapsing over table factors	51
2.7.2	Collapsing table levels	53
2.8	Converting among frequency tables and data frames	53
2.8.1	Table form to frequency form	54
2.8.2	Case form to table form	55
2.8.3	Table form to case form	55
2.8.4	Publishing tables to L ^A T _E X or HTML	56
2.9	A complex example: TV viewing data*	58
2.9.1	Creating data frames and arrays	58
2.9.2	Subsetting and collapsing	60
2.10	Lab exercises	60
3	Fitting and Graphing Discrete Distributions	65
3.1	Introduction to discrete distributions	66
3.1.1	Binomial data	66
3.1.2	Poisson data	69
3.1.3	Type-token distributions	72
3.2	Characteristics of discrete distributions	73
3.2.1	The binomial distribution	74
3.2.2	The Poisson distribution	76
3.2.3	The negative binomial distribution	82
3.2.4	The geometric distribution	85
3.2.5	The logarithmic series distribution	86
3.2.6	Power series family	86
3.3	Fitting discrete distributions	87
3.3.1	R tools for discrete distributions	89
3.3.2	Plots of observed and fitted frequencies	92
3.4	Diagnosing discrete distributions: Ord plots	95
3.5	Poissonness plots and generalized distribution plots	99
3.5.1	Features of the Poissonness plot	100
3.5.2	Plot construction	100
3.5.3	The distplot function	101
3.5.4	Plots for other distributions	102
3.6	Fitting discrete distributions as generalized linear models*	104
3.6.1	Covariates, overdispersion, and excess zeros	107
3.7	Chapter summary	109

3.8	Lab exercises	109
II Exploratory and Hypothesis-Testing Methods		113
4	Two-Way Contingency Tables	115
4.1	Introduction	115
4.2	Tests of association for two-way tables	119
4.2.1	Notation and terminology	119
4.2.2	2 by 2 tables: Odds and odds ratios	121
4.2.3	Larger tables: Overall analysis	124
4.2.4	Tests for ordinal variables	125
4.2.5	Sample CMH profiles	126
4.3	Stratified analysis	127
4.3.1	Computing strata-wise statistics	128
4.3.2	Assessing homogeneity of association	129
4.4	Fourfold display for 2 x 2 tables	130
4.4.1	Confidence rings for odds ratio	133
4.4.2	Stratified analysis for 2 x 2 x k tables	133
4.5	Sieve diagrams	138
4.5.1	Two-way tables	138
4.5.2	Larger tables: The strucplot framework	141
4.6	Association plots	145
4.7	Observer agreement	146
4.7.1	Measuring agreement	148
4.7.2	Observer agreement chart	150
4.7.3	Observer bias in agreement	152
4.8	Trilinear plots	153
4.9	Chapter summary	157
4.10	Lab exercises	158
5	Mosaic Displays for n-Way Tables	161
5.1	Introduction	161
5.2	Two-way tables	162
5.2.1	Shading levels	166
5.2.2	Interpretation and reordering	166
5.3	The strucplot framework	167
5.3.1	Components overview	167
5.3.2	Shading schemes	169
5.4	Three-way and larger tables	176
5.4.1	A primer on loglinear models	177
5.4.2	Fitting models	179
5.5	Model and plot collections	183
5.5.1	Sequential plots and models	184
5.5.2	Causal models	186
5.5.3	Partial association	188
5.6	Mosaic matrices for categorical data	197
5.6.1	Mosaic matrices for pairwise associations	197
5.6.2	Generalized mosaic matrices and pairs plots	201
5.7	3D mosaics	203
5.8	Visualizing the structure of loglinear models	205
5.8.1	Mutual independence	206

5.8.2	Joint independence	208
5.9	Related visualization methods	209
5.9.1	Doubled-decker plots	209
5.9.2	Generalized odds ratios*	211
5.10	Chapter summary	215
5.11	Lab exercises	216
6	Correspondence Analysis	221
6.1	Introduction	221
6.2	Simple correspondence analysis	222
6.2.1	Notation and terminology	222
6.2.2	Geometric and statistical properties	224
6.2.3	R software for correspondence analysis	224
6.2.4	Correspondence analysis and mosaic displays	231
6.3	Multi-way tables: Stacking and other tricks	232
6.3.1	Interactive coding in R	233
6.3.2	Marginal tables and supplementary variables	238
6.4	Multiple correspondence analysis	240
6.4.1	Bivariate MCA	240
6.4.2	The Burt matrix	243
6.4.3	Multivariate MCA	243
6.5	Biplots for contingency tables	248
6.5.1	CA bilinear biplots	248
6.5.2	Biadditive biplots	252
6.6	Chapter summary	254
6.7	Lab exercises	254
III	Model-Building Methods	259
7	Logistic Regression Models	261
7.1	Introduction	261
7.2	The logistic regression model	263
7.2.1	Fitting a logistic regression model	265
7.2.2	Model tests for simple logistic regression	267
7.2.3	Plotting a binary response	268
7.2.4	Grouped binomial data	270
7.3	Multiple logistic regression models	272
7.3.1	Conditional plots	275
7.3.2	Full-model plots	276
7.3.3	Effect plots	278
7.4	Case studies	281
7.4.1	Simple models: Group comparisons and effect plots	282
7.4.2	More complex models: Model selection and visualization	294
7.5	Influence and diagnostic plots	303
7.5.1	Residuals and leverage	303
7.5.2	Influence diagnostics	304
7.5.3	Other diagnostic plots*	312
7.6	Chapter summary	319
7.7	Lab exercises	320

8 Models for Polytomous Responses	323
8.1 Ordinal response	324
8.1.1 Latent variable interpretation	325
8.1.2 Fitting the proportional odds model	326
8.1.3 Testing the proportional odds assumption	327
8.1.4 Graphical assessment of proportional odds	329
8.1.5 Visualizing results for the proportional odds model	331
8.2 Nested dichotomies	335
8.3 Generalized logit model	341
8.4 Chapter summary	346
8.5 Lab exercises	346
9 Loglinear and Logit Models for Contingency Tables	349
9.1 Introduction	349
9.2 Loglinear models for frequencies	350
9.2.1 Loglinear models as ANOVA models for frequencies	350
9.2.2 Loglinear models for three-way tables	352
9.2.3 Loglinear models as GLMs for frequencies	352
9.3 Fitting and testing loglinear models	353
9.3.1 Model fitting functions	353
9.3.2 Goodness-of-fit tests	354
9.3.3 Residuals for loglinear models	356
9.3.4 Using loglm()	357
9.3.5 Using glm()	359
9.4 Equivalent logit models	363
9.5 Zero frequencies	368
9.6 Chapter summary	372
9.7 Lab exercises	372
10 Extending Loglinear Models	375
10.1 Models for ordinal variables	376
10.1.1 Loglinear models for ordinal variables	376
10.1.2 Visualizing model structure	381
10.1.3 Log-multiplicative (RC) models	382
10.2 Square tables	389
10.2.1 Quasi-independence, symmetry, quasi-symmetry, and topological models	389
10.2.2 Ordinal square tables	396
10.3 Three-way and higher-dimensional tables	400
10.4 Multivariate responses*	403
10.4.1 Bivariate, binary response models	405
10.4.2 More complex models	415
10.5 Chapter summary	425
10.6 Lab exercises	426
11 Generalized Linear Models for Count Data	429
11.1 Components of generalized linear models	430
11.1.1 Variance functions	431
11.1.2 Hypothesis tests for coefficients	432
11.1.3 Goodness-of-fit tests	433
11.1.4 Comparing non-nested models	434

xii	Contents
11.2 GLMs for count data	435
11.3 Models for overdispersed count data	444
11.3.1 The quasi-Poisson model	445
11.3.2 The negative-binomial model	446
11.3.3 Visualizing the mean–variance relation	447
11.3.4 Testing overdispersion	449
11.3.5 Visualizing goodness-of-fit	450
11.4 Models for excess zero counts	451
11.4.1 Zero-inflated models	452
11.4.2 Hurdle models	454
11.4.3 Visualizing zero counts	454
11.5 Case studies	456
11.5.1 Cod parasites	456
11.5.2 Demand for medical care by the elderly	468
11.6 Diagnostic plots for model checking	480
11.6.1 Diagnostic measures and residuals for GLMs	480
11.6.2 Quantile–quantile and half-normal plots	485
11.7 Multivariate response GLM models*	489
11.7.1 Analyzing correlations: HE plots	491
11.7.2 Analyzing associations: Odds ratios and fourfold plots	492
11.8 Chapter summary	500
11.9 Lab exercises	501
References	505
Author Index	525
Example Index	529
Subject Index	531