Contents

List of Figures xvi
List of Tables xxi
Preface xxiii
Acknowledgments xxv

Part I THE PRELIMINARIES

1 Why R? 3
1.1 Why R? 3
1.2 R Installation 5
1.3 There is Nothing such as PRACTICALS 5
1.4 Datasets in R and Internet 6
1.4.1 List of Web-sites containing DATASETS 7
1.4.2 Antique Datasets 8
1.5 http://cran.r-project.org 9
1.5.1 http://r-project.org 10
1.5.2 http://www.cran.r-project.org/web/views/ 10
1.5.3 Is subscribing to R-Mailing List useful? 10
1.6 R and its Interface with other Software 11
1.7 help and/or? 11
1.8 R Books 12
1.9 A Road Map 13

2 The R Basics 15
2.1 Introduction 15
2.2 Simple Arithmetics and a Little Beyond 16
2.2.1 Absolute Values, Remainders, etc. 16
2.2.2 round, floor, etc. 17
2.2.3 Summary Functions 18
2.2.4 Trigonometric Functions 18
2.2.5 Complex Numbers 19
2.2.6 Special Mathematical Functions 21

2.3 Some Basic R Functions 22
2.3.1 Summary Statistics 23
2.3.2 `is`, `as`, `is.na`, etc. 25
2.3.3 `factors`, `levels`, etc. 26
2.3.4 Control Programming 27
2.3.5 Other Useful Functions 29
2.3.6 Calculus* 31

2.4 Vectors and Matrices in R 33
2.4.1 Vectors 33
2.4.2 Matrices 36

2.5 Data Entering and Reading from Files 41
2.5.1 Data Entering 41
2.5.2 Reading Data from External Files 43

2.6 Working with Packages 44
2.7 R Session Management 45
2.8 Further Reading 46
2.9 Complements, Problems, and Programs 46

3 Data Preparation and Other Tricks 49
3.1 Introduction 49
3.2 Manipulation with Complex Format Files 50
3.3 Reading Datasets of Foreign Formats 55
3.4 Displaying R Objects 56
3.5 Manipulation Using R Functions 57
3.6 Working with Time and Date 59
3.7 Text Manipulations 62
3.8 Scripts and Text Editors for R 64
3.8.1 Text Editors for Linuxians 64
3.9 Further Reading 65
3.10 Complements, Problems, and Programs 65

4 Exploratory Data Analysis 67
4.1 Introduction: The Tukey’s School of Statistics 67
4.2 Essential Summaries of EDA 68
4.3 Graphical Techniques in EDA 71
4.3.1 Boxplot 71
4.3.2 Histogram 76
4.3.3 Histogram Extensions and the Rootogram 79
4.3.4 Pareto Chart 81
4.3.5 Stem-and-Leaf Plot 84
4.3.6 Run Chart 88
4.3.7 Scatter Plot 89
4.4 Quantitative Techniques in EDA 91
   4.4.1 Trimean 91
   4.4.2 Letter Values 92
4.5 Exploratory Regression Models 95
   4.5.1 Resistant Line 95
   4.5.2 Median Polish 98
4.6 Further Reading 99
4.7 Complements, Problems, and Programs 100

Part II  PROBABILITY AND INFEERENCE

5 Probability Theory 105
5.1 Introduction 105
5.2 Sample Space, Set Algebra, and Elementary Probability 106
5.3 Counting Methods
   5.3.1 Sampling: The Diverse Ways 114
   5.3.2 The Binomial Coefficients and the Pascals Triangle 118
   5.3.3 Some Problems Based on Combinatorics 119
5.4 Probability: A Definition 122
   5.4.1 The Prerequisites 122
   5.4.2 The Kolmogorov Definition 127
5.5 Conditional Probability and Independence 130
5.6 Bayes Formula 132
5.7 Random Variables, Expectations, and Moments
   5.7.1 The Definition 133
   5.7.2 Expectation of Random Variables 136
5.8 Distribution Function, Characteristic Function, and Moment
   Generation Function 143
5.9 Inequalities
   5.9.1 The Markov Inequality 145
   5.9.2 The Jensen's Inequality 145
   5.9.3 The Chebyshev Inequality 146
5.10 Convergence of Random Variables 146
   5.10.1 Convergence in Distributions 147
   5.10.2 Convergence in Probability 150
   5.10.3 Convergence in \( r^{th} \) Mean 150
   5.10.4 Almost Sure Convergence 151
5.11 The Law of Large Numbers 152
   5.11.1 The Weak Law of Large Numbers 152
5.12 The Central Limit Theorem
   5.12.1 The de Moivre-Laplace Central Limit Theorem 153
   5.12.2 CLT for iid Case 154
   5.12.3 The Lindeberg-Feller CLT 157
   5.12.4 The Liapounou CLT 162
5.13 Further Reading 165
5.13.1 Intuitive, Elementary, and First Course Source 165
5.13.2 The Classics and Second Course Source 166
5.13.3 The Problem Books 167
5.13.4 Other Useful Sources 167
5.13.5 R for Probability 167
5.14 Complements, Problems, and Programs 167

6 Probability and Sampling Distributions 171
6.1 Introduction 171
6.2 Discrete Univariate Distributions 172
  6.2.1 The Discrete Uniform Distribution 172
  6.2.2 The Binomial Distribution 173
  6.2.3 The Geometric Distribution 176
  6.2.4 The Negative Binomial Distribution 178
  6.2.5 Poisson Distribution 179
  6.2.6 The Hypergeometric Distribution 182
6.3 Continuous Univariate Distributions 184
  6.3.1 The Uniform Distribution 184
  6.3.2 The Beta Distribution 186
  6.3.3 The Exponential Distribution 187
  6.3.4 The Gamma Distribution 188
  6.3.5 The Normal Distribution 189
  6.3.6 The Cauchy Distribution 191
  6.3.7 The t-Distribution 193
  6.3.8 The Chi-square Distribution 193
  6.3.9 The F-Distribution 194
6.4 Multivariate Probability Distributions 194
  6.4.1 The Multinomial Distribution 194
  6.4.2 Dirichlet Distribution 195
  6.4.3 The Multivariate Normal Distribution 195
  6.4.4 The Multivariate t Distribution 196
6.5 Populations and Samples 196
6.6 Sampling from the Normal Distributions 197
6.7 Some Finer Aspects of Sampling Distributions 201
  6.7.1 Sampling Distribution of Median 201
  6.7.2 Sampling Distribution of Mean of Standard Distributions 201
6.8 Multivariate Sampling Distributions 203
  6.8.1 Noncentral Univariate Chi-square, t, and F Distributions 203
  6.8.2 Wishart Distribution 205
  6.8.3 Hotellings $T^2$ Distribution 206
6.9 Bayesian Sampling Distributions 206
6.10 Further Reading 207
6.11 Complements, Problems, and Programs 208

7 Parametric Inference 209
7.1 Introduction 209
## Contents

7.2 Families of Distribution
7.2.1 *The Exponential Family* 210
7.2.2 Pitman Family 213
7.3 Loss Functions 214
7.4 Data Reduction
7.4.1 Sufficiency 217
7.4.2 Minimal Sufficiency 219
7.5 Likelihood and Information
7.5.1 *The Likelihood Principle* 220
7.5.2 *The Fisher Information* 226
7.6 Point Estimation
7.6.1 *Maximum Likelihood Estimation* 231
7.6.2 Method of Moments Estimator 239
7.7 Comparison of Estimators
7.7.1 *Unbiased Estimators* 241
7.7.2 Improving Unbiased Estimators 243
7.8 Confidence Intervals 245
7.9 Testing Statistical Hypotheses–The Preliminaries 246
7.10 The Neyman-Pearson Lemma 251
7.11 Uniformly Most Powerful Tests 256
7.12 Uniformly Most Powerful Unbiased Tests
7.12.1 Tests for the Means: One- and Two-Sample *t*-Test 263
7.13 Likelihood Ratio Tests
7.13.1 *Normal Distribution: One-Sample Problems* 266
7.13.2 *Normal Distribution: Two-Sample Problem for the Mean* 269
7.14 Behrens-Fisher Problem 270
7.15 Multiple Comparison Tests
7.15.1 *Bonferroni’s Method* 272
7.15.2 *Holm’s Method* 273
7.16 The EM Algorithm*
7.16.1 Introduction 274
7.16.2 The Algorithm 274
7.16.3 Introductory Applications 275
7.17 Further Reading
7.17.1 Early Classics 280
7.17.2 *Texts from the Last 30 Years* 281
7.18 Complements, Problems, and Programs 281

8 Nonparametric Inference

8.1 Introduction 283
8.2 Empirical Distribution Function and Its Applications
8.2.1 Statistical Functionals 285
8.3 The Jackknife and Bootstrap Methods
8.3.1 The Jackknife 288
8.3.2 The Bootstrap 289
8.3.3 Bootstrapping Simple Linear Model* 292
8.4 Non-parametric Smoothing 294
  8.4.1 Histogram Smoothing 294
  8.4.2 Kernel Smoothing 297
  8.4.3 Nonparametric Regression Models* 300
8.5 Non-parametric Tests 304
  8.5.1 The Wilcoxon Signed-Ranks Test 305
  8.5.2 The Mann-Whitney test 308
  8.5.3 The Siegel-Tukey Test 309
  8.5.4 The Wald-Wolfowitz Run Test 311
  8.5.5 The Kolmogorov-Smirnov Test 312
  8.5.6 Kruskal-Wallis Test* 314
8.6 Further Reading 315
8.7 Complements, Problems, and Programs 316

9 Bayesian Inference 317
  9.1 Introduction 317
  9.2 Bayesian Probabilities 317
  9.3 The Bayesian Paradigm for Statistical Inference 321
    9.3.1 Bayesian Sufficiency and the Principle 321
    9.3.2 Bayesian Analysis and Likelihood Principle 322
    9.3.3 Informative and Conjugate Prior 322
    9.3.4 Non-informative Prior 323
  9.4 Bayesian Estimation 323
    9.4.1 Inference for Binomial Distribution 323
    9.4.2 Inference for the Poisson Distribution 326
    9.4.3 Inference for Uniform Distribution 327
    9.4.4 Inference for Exponential Distribution 328
    9.4.5 Inference for Normal Distributions 329
  9.5 The Credible Intervals 332
  9.6 Bayes Factors for Testing Problems 333
  9.7 Further Reading 334
  9.8 Complements, Problems, and Programs 335

Part III STOCHASTIC PROCESSES AND MONTE CARLO

10 Stochastic Processes 339
  10.1 Introduction 339
  10.2 Kolmogorov’s Consistency Theorem 340
  10.3 Markov Chains 341
    10.3.1 The m-Step TPM 344
    10.3.2 Classification of States 345
    10.3.3 Canonical Decomposition of an Absorbing Markov Chain 347
    10.3.4 Stationary Distribution and Mean First Passage Time of an
              Ergodic Markov Chain 350
    10.3.5 Time Reversible Markov Chain 352
10.4 Application of Markov Chains in Computational Statistics 352
  10.4.1 The Metropolis-Hastings Algorithm 353
  10.4.2 Gibbs Sampler 354
  10.4.3 Illustrative Examples 355
10.5 Further Reading 361
10.6 Complements, Problems, and Programs 361

11 Monte Carlo Computations 363
11.1 Introduction 363
11.2 Generating the (Pseudo-) Random Numbers 364
  11.2.1 Useful Random Generators 364
  11.2.2 Probability Through Simulation 366
11.3 Simulation from Probability Distributions and Some Limit Theorems 373
  11.3.1 Simulation from Discrete Distributions 373
  11.3.2 Simulation from Continuous Distributions 380
  11.3.3 Understanding Limit Theorems through Simulation 383
  11.3.4 Understanding The Central Limit Theorem 386
11.4 Monte Carlo Integration 388
11.5 The Accept-Reject Technique 390
11.6 Application to Bayesian Inference 394
11.7 Further Reading 397
11.8 Complements, Problems, and Programs 397

Part IV LINEAR MODELS

12 Linear Regression Models 401
12.1 Introduction 401
12.2 Simple Linear Regression Model 402
  12.2.1 Fitting a Linear Model 403
  12.2.2 Confidence Intervals 405
  12.2.3 The Analysis of Variance (ANOVA) 407
  12.2.4 The Coefficient of Determination 409
  12.2.5 The “lm” Function from R 410
  12.2.6 Residuals for Validation of the Model Assumptions 412
  12.2.7 Prediction for the Simple Regression Model 416
  12.2.8 Regression through the Origin 417
12.3 The Anscombe Warnings and Regression Abuse 418
12.4 Multiple Linear Regression Model 421
  12.4.1 Scatter Plots: A First Look 422
  12.4.2 Other Useful Graphical Methods 423
  12.4.3 Fitting a Multiple Linear Regression Model 427
  12.4.4 Testing Hypotheses and Confidence Intervals 429
12.5 Model Diagnostics for the Multiple Regression Model 433
  12.5.1 Residuals 433
  12.5.2 Influence and Leverage Diagnostics 436
12.6 Multicollinearity
  12.6.1 Variance Inflation Factor
  12.6.2 Eigen System Analysis
12.7 Data Transformations
  12.7.1 Linearization
  12.7.2 Variance Stabilization
  12.7.3 Power Transformation
12.8 Model Selection
  12.8.1 Backward Elimination
  12.8.2 Forward and Stepwise Selection
12.9 Further Reading
  12.9.1 Early Classics
  12.9.2 Industrial Applications
  12.9.3 Regression Details
  12.9.4 Modern Regression Texts
  12.9.5 R for Regression
12.10 Complements, Problems, and Programs

13 Experimental Designs
13.1 Introduction
13.2 Principles of Experimental Design
13.3 Completely Randomized Designs
  13.3.1 The CRD Model
  13.3.2 Randomization in CRD
  13.3.3 Inference for the CRD Models
  13.3.4 Validation of Model Assumptions
  13.3.5 Contrasts and Multiple Testing for the CRD Model
13.4 Block Designs
  13.4.1 Randomization and Analysis of Balanced Block Designs
  13.4.2 Incomplete Block Designs
  13.4.3 Latin Square Design
  13.4.4 Graeco Latin Square Design
13.5 Factorial Designs
  13.5.1 Two Factorial Experiment
  13.5.2 Three-Factorial Experiment
  13.5.3 Blocking in Factorial Experiments
13.6 Further Reading
13.7 Complements, Problems, and Programs

14 Multivariate Statistical Analysis - I
14.1 Introduction
14.2 Graphical Plots for Multivariate Data
14.3 Definitions, Notations, and Summary Statistics for Multivariate Data
  14.3.1 Definitions and Data Visualization
  14.3.2 Early Outlier Detection
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.4</td>
<td>Testing for Mean Vectors : One Sample</td>
<td>520</td>
</tr>
<tr>
<td>14.4.1</td>
<td>Testing for Mean Vector with Known Variance-Covariance Matrix</td>
<td>520</td>
</tr>
<tr>
<td>14.4.2</td>
<td>Testing for Mean Vectors with Unknown Variance-Covariance Matrix</td>
<td>521</td>
</tr>
<tr>
<td>14.5</td>
<td>Testing for Mean Vectors : Two-Samples</td>
<td>523</td>
</tr>
<tr>
<td>14.6</td>
<td>Multivariate Analysis of Variance</td>
<td>526</td>
</tr>
<tr>
<td>14.6.1</td>
<td>Wilks Test Statistic</td>
<td>526</td>
</tr>
<tr>
<td>14.6.2</td>
<td>Roy's Test</td>
<td>528</td>
</tr>
<tr>
<td>14.6.3</td>
<td>Pillai's Test Statistic</td>
<td>529</td>
</tr>
<tr>
<td>14.6.4</td>
<td>The Lawley-Hotelling Test Statistic</td>
<td>529</td>
</tr>
<tr>
<td>14.7</td>
<td>Testing for Variance-Covariance Matrix: One Sample</td>
<td>531</td>
</tr>
<tr>
<td>14.7.1</td>
<td>Testing for Sphericity</td>
<td>532</td>
</tr>
<tr>
<td>14.8</td>
<td>Testing for Variance-Covariance Matrix: k-Samples</td>
<td>533</td>
</tr>
<tr>
<td>14.9</td>
<td>Testing for Independence of Sub-vectors</td>
<td>536</td>
</tr>
<tr>
<td>14.10</td>
<td>Further Reading</td>
<td>538</td>
</tr>
<tr>
<td>14.11</td>
<td>Complements, Problems, and Programs</td>
<td>538</td>
</tr>
<tr>
<td>15</td>
<td>Multivariate Statistical Analysis - II</td>
<td>541</td>
</tr>
<tr>
<td>15.1</td>
<td>Introduction</td>
<td>541</td>
</tr>
<tr>
<td>15.2</td>
<td>Classification and Discriminant Analysis</td>
<td>541</td>
</tr>
<tr>
<td>15.2.1</td>
<td>Discrimination Analysis</td>
<td>542</td>
</tr>
<tr>
<td>15.2.2</td>
<td>Classification</td>
<td>543</td>
</tr>
<tr>
<td>15.3</td>
<td>Canonical Correlations</td>
<td>544</td>
</tr>
<tr>
<td>15.4</td>
<td>Principal Component Analysis – Theory and Illustration</td>
<td>547</td>
</tr>
<tr>
<td>15.4.1</td>
<td>The Theory</td>
<td>547</td>
</tr>
<tr>
<td>15.4.2</td>
<td>Illustration Through a Dataset</td>
<td>549</td>
</tr>
<tr>
<td>15.5</td>
<td>Applications of Principal Component Analysis</td>
<td>553</td>
</tr>
<tr>
<td>15.5.1</td>
<td>PCA for Linear Regression</td>
<td>553</td>
</tr>
<tr>
<td>15.5.2</td>
<td>Biplots</td>
<td>556</td>
</tr>
<tr>
<td>15.6</td>
<td>Factor Analysis</td>
<td>560</td>
</tr>
<tr>
<td>15.6.1</td>
<td>The Orthogonal Factor Analysis Model</td>
<td>561</td>
</tr>
<tr>
<td>15.6.2</td>
<td>Estimation of Loadings and Communalities</td>
<td>562</td>
</tr>
<tr>
<td>15.7</td>
<td>Further Reading</td>
<td>568</td>
</tr>
<tr>
<td>15.7.1</td>
<td>The Classics and Applied Perspectives</td>
<td>568</td>
</tr>
<tr>
<td>15.7.2</td>
<td>Multivariate Analysis and Software</td>
<td>568</td>
</tr>
<tr>
<td>15.8</td>
<td>Complements, Problems, and Programs</td>
<td>569</td>
</tr>
<tr>
<td>16</td>
<td>Categorical Data Analysis</td>
<td>571</td>
</tr>
<tr>
<td>16.1</td>
<td>Introduction</td>
<td>571</td>
</tr>
<tr>
<td>16.2</td>
<td>Graphical Methods for CDA</td>
<td>572</td>
</tr>
<tr>
<td>16.2.1</td>
<td>Bar and Stacked Bar Plots</td>
<td>572</td>
</tr>
<tr>
<td>16.2.2</td>
<td>Spine Plots</td>
<td>575</td>
</tr>
<tr>
<td>16.2.3</td>
<td>Mosaic Plots</td>
<td>577</td>
</tr>
<tr>
<td>16.2.4</td>
<td>Pie Charts and Dot Charts</td>
<td>580</td>
</tr>
<tr>
<td>16.2.5</td>
<td>Four-Fold Plots</td>
<td>583</td>
</tr>
<tr>
<td>16.3</td>
<td>The Odds Ratio</td>
<td>586</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>16.4</td>
<td>The Simpson’s Paradox</td>
<td>588</td>
</tr>
<tr>
<td>16.5</td>
<td>The Binomial, Multinomial, and Poisson Models</td>
<td>589</td>
</tr>
<tr>
<td>16.5.1</td>
<td>The Binomial Model</td>
<td>589</td>
</tr>
<tr>
<td>16.5.2</td>
<td>The Multinomial Model</td>
<td>590</td>
</tr>
<tr>
<td>16.5.3</td>
<td>The Poisson Model</td>
<td>591</td>
</tr>
<tr>
<td>16.6</td>
<td>The Problem of Overdispersion</td>
<td>593</td>
</tr>
<tr>
<td>16.7</td>
<td>The $\chi^2$-Tests of Independence</td>
<td>593</td>
</tr>
<tr>
<td>16.8</td>
<td>Further Reading</td>
<td>595</td>
</tr>
<tr>
<td>16.9</td>
<td>Complements, Problems, and Programs</td>
<td>595</td>
</tr>
<tr>
<td>17</td>
<td><strong>Generalized Linear Models</strong></td>
<td>597</td>
</tr>
<tr>
<td>17.1</td>
<td>Introduction</td>
<td>597</td>
</tr>
<tr>
<td>17.2</td>
<td>Regression Problems in Count/Discrete Data</td>
<td>597</td>
</tr>
<tr>
<td>17.3</td>
<td>Exponential Family and the GLM</td>
<td>600</td>
</tr>
<tr>
<td>17.4</td>
<td>The Logistic Regression Model</td>
<td>601</td>
</tr>
<tr>
<td>17.5</td>
<td>Inference for the Logistic Regression Model</td>
<td>602</td>
</tr>
<tr>
<td>17.5.1</td>
<td>Estimation of the Regression Coefficients and Related Parameters</td>
<td>602</td>
</tr>
<tr>
<td>17.5.2</td>
<td>Estimation of the Variance-Covariance Matrix of $\hat{\beta}$</td>
<td>606</td>
</tr>
<tr>
<td>17.5.3</td>
<td>Confidence Intervals and Hypotheses Testing for the Regression Coefficients</td>
<td>607</td>
</tr>
<tr>
<td>17.5.4</td>
<td>Residuals for the Logistic Regression Model</td>
<td>608</td>
</tr>
<tr>
<td>17.5.5</td>
<td>Deviance Test and Hosmer-Lemeshow Goodness-of-Fit Test</td>
<td>611</td>
</tr>
<tr>
<td>17.6</td>
<td>Model Selection in Logistic Regression Models</td>
<td>613</td>
</tr>
<tr>
<td>17.7</td>
<td>Probit Regression</td>
<td>618</td>
</tr>
<tr>
<td>17.8</td>
<td>Poisson Regression Model</td>
<td>621</td>
</tr>
<tr>
<td>17.9</td>
<td>Further Reading</td>
<td>625</td>
</tr>
<tr>
<td>17.10</td>
<td>Complements, Problems, and Programs</td>
<td>626</td>
</tr>
<tr>
<td>Appendix A</td>
<td>Open Source Software—An Epilogue</td>
<td>627</td>
</tr>
<tr>
<td>Appendix B</td>
<td>The Statistical Tables</td>
<td>631</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>633</td>
</tr>
<tr>
<td>Author Index</td>
<td></td>
<td>643</td>
</tr>
<tr>
<td>Subject Index</td>
<td></td>
<td>649</td>
</tr>
<tr>
<td>R Codes</td>
<td></td>
<td>659</td>
</tr>
</tbody>
</table>