Contents

Preface to the second edition — V

1 Introduction — 1
1.1 Relationships between mathematical subjects and data science — 2
1.2 Structure of the book — 4
1.2.1 Part one — 4
1.2.2 Part two — 4
1.2.3 Part three — 5
1.3 Our motivation for writing this book — 5
1.4 Examples and listings — 6
1.5 How to use this book — 7

Part I: Introduction to R

2 Overview of programming paradigms — 11
2.1 Introduction — 11
2.2 Imperative programming — 12
2.3 Functional programming — 13
2.4 Object-oriented programming — 15
2.5 Logic programming — 17
2.6 Other programming paradigms — 18
2.6.1 The multiparadigm language R — 18
2.7 Compiler versus interpreter languages — 20
2.8 Semantics of programming languages — 21
2.9 Further reading — 21
2.10 Summary — 22

3 Setting up and installing the R program — 23
3.1 Installing R on Linux — 23
3.2 Installing R on MAC OS X — 24
3.3 Installing R on Windows — 24
3.4 Using R — 24
3.5 Summary — 24

4 Installation of R packages — 26
4.1 Installing packages from CRAN — 26
4.2 Installing packages from Bioconductor — 26
4.3 Installing packages from GitHub — 27
4.4 Installing packages manually — 27
4.4.1 Terminal and unix commands — 27
4.4.2 Package installation — 28
4.5 Activation of a package in an R session — 29
4.6 Summary — 29

5 Introduction to programming in R — 30
5.1 Basic elements of R — 30
5.1.1 Navigating directories — 31
5.1.2 System functions — 31
5.1.3 Getting help — 32
5.2 Basic programming — 33
5.2.1 If-clause — 33
5.2.2 Switch — 34
5.2.3 Loops — 35
5.2.4 For-loop — 35
5.2.5 While-loop — 35
5.2.6 Logic behind a For-loop — 36
5.2.7 Break — 39
5.2.8 Repeat-loop — 39
5.3 Data structures — 39
5.3.1 Vector — 39
5.3.2 Matrix — 42
5.3.3 List — 45
5.3.4 Array — 46
5.3.5 Data frame — 47
5.3.6 Environment — 48
5.3.7 Removing variables from the workspace — 49
5.3.8 Factor — 49
5.3.9 Date and Time — 50
5.3.10 Information about R objects — 50
5.4 Handling character strings — 51
5.4.1 The function nchar() — 51
5.4.2 The function paste() — 52
5.4.3 The function substr() — 52
5.4.4 The function strsplit() — 53
5.4.5 Regular expressions — 53
5.5 Sorting vectors — 56
5.6 Writing functions — 57
5.6.1 One input argument and one output value — 57
5.6.2 Scope of variables — 59
5.6.3 One input argument, many output values — 60
5.6.4 Many input arguments, many output values — 61
5.7 Writing and reading data — 61
  5.7.1 Writing data to a file — 61
  5.7.2 Reading data from a file — 63
  5.7.3 Low level reading functions — 64
  5.7.4 Summary of writing and reading functions — 67
  5.7.5 Other data formats — 67
5.8 Useful commands — 68
  5.8.1 The function which() — 68
  5.8.2 The function apply() — 69
  5.8.3 Set commands — 70
  5.8.4 The function unique() — 70
  5.8.5 Testing arguments and converting variables — 71
  5.8.6 The function sample() — 71
  5.8.7 The function try() — 73
  5.8.8 The function system() — 74
5.9 Practical usage of R — 74
  5.9.1 Advantage over GUI software — 75
5.10 Summary — 75

6 Creating R packages — 76
  6.1 Requirements — 76
    6.1.1 R base packages — 76
    6.1.2 R repositories — 77
    6.1.3 Rtools — 77
  6.2 R code optimization — 77
    6.2.1 Profiling an R script — 78
    6.2.2 Byte code compilation — 78
    6.2.3 GPU library, code, and others — 79
    6.2.4 Exception handling — 79
  6.3 S3, S4, and RC object-oriented systems — 80
    6.3.1 The S3 class — 80
    6.3.2 The S4 class — 82
    6.3.3 Reference class (RC) system — 83
  6.4 Creating an R package based on the S3 class system — 84
    6.4.1 R program file — 84
    6.4.2 Building an R package — 86
  6.5 Checking the package — 87
  6.6 Installation and usage of the package — 87
  6.7 Loading and using a package — 88
    6.7.1 Content of the files edited when generating the package — 88
  6.8 Summary — 91
Part II: **Graphics in R**

7 **Basic plotting functions** — 95
7.1 Plot — 95
7.1.1 Adding multiple curves in one plot — 97
7.1.2 Adding horizontal and vertical lines — 99
7.1.3 Opening a new figure window — 100
7.2 Histograms — 100
7.3 Bar plots — 101
7.4 Pie charts — 102
7.5 Dot plots — 102
7.6 Strip and rug plots — 105
7.7 Density plots — 107
7.8 Combining a scatterplot with histograms: the layout function — 110
7.9 Three-dimensional plots — 111
7.10 Contour and image plots — 112
7.11 Summary — 113

8 **Advanced plotting functions: ggplot2** — 114
8.1 Introduction — 114
8.2 `qplot()` — 114
8.3 `ggplot()` — 117
8.3.1 Simple examples — 118
8.3.2 Multiple data sets — 120
8.3.3 `geoms()` — 122
8.3.4 Smoothing — 125
8.4 Summary — 128

9 **Visualization of networks** — 129
9.1 Introduction — 129
9.2 igraph — 129
9.2.1 Generation of regular and complex networks — 131
9.2.2 Basic network attributes — 132
9.2.3 Layout styles — 135
9.2.4 Plotting networks — 136
9.2.5 Analyzing and manipulating networks — 137
9.3 NetBioV — 137
9.3.1 Global network layout — 138
9.3.2 Modular network layout — 138
9.3.3 Layered network (multiroot) layout — 140
9.3.4 Further features — 140
9.3.5 Examples: Visualization of networks using NetBioV — 141
Part III: Mathematical basics of data science

10 Mathematics as a language for science — 149
  10.1 Introduction — 149
  10.2 Numbers and number operations — 151
  10.3 Sets and set operations — 153
  10.4 Boolean logic — 155
  10.5 Sum, product, and Binomial coefficients — 157
  10.6 Further symbols — 161
  10.7 Importance of definitions and theorems — 164
  10.8 Summary — 165

11 Computability and complexity — 166
  11.1 Introduction — 166
  11.2 A brief history of computer science — 167
  11.3 Turing machines — 168
  11.4 Computability — 169
  11.5 Complexity of algorithms — 170
    11.5.1 Bounds — 171
    11.5.2 Examples — 172
    11.5.3 Important properties of the $O$-notation — 174
  11.6 Known complexity classes — 174
  11.7 Summary — 175

12 Linear algebra — 176
  12.1 Vectors and matrices — 176
    12.1.1 Vectors — 176
    12.1.2 Vector representations in other coordinates systems — 189
    12.1.3 Matrices — 197
  12.2 Operations with matrices — 201
  12.3 Special matrices — 204
  12.4 Trace and determinant of a matrix — 205
  12.5 Subspaces, dimension, and rank of a matrix — 206
  12.6 Eigenvalues and eigenvectors of a matrix — 209
  12.7 Matrix norms — 212
  12.8 Matrix factorization — 213
    12.8.1 LU factorization — 213
    12.8.2 Cholesky factorization — 215
    12.8.3 QR factorization — 216
12.8.4 Singular value decomposition — 218
12.9 Systems of linear equations — 221
12.10 Exercises — 224

13 Analysis — 225
13.1 Introduction — 225
13.2 Limiting values — 225
13.3 Differentiation — 228
13.4 Extrema of a function — 233
13.5 Taylor series expansion — 235
13.6 Integrals — 239
13.6.1 Properties of definite integrals — 240
13.6.2 Numerical integration — 240
13.7 Polynomial interpolation — 241
13.8 Root finding methods — 243
13.9 Further reading — 247
13.10 Exercises — 247

14 Differential equations — 249
14.1 Ordinary differential equations (ODE) — 249
14.1.1 Initial value ODE problems — 249
14.2 Partial differential equations (PDE) — 254
14.2.1 First-order PDE — 255
14.2.2 Second-order PDE — 255
14.2.3 Boundary and initial conditions — 256
14.2.4 Well-posed PDE problems — 256
14.3 Exercises — 260

15 Dynamical systems — 262
15.1 Introduction — 262
15.2 Population growth models — 264
15.2.1 Exponential population growth model — 264
15.2.2 Logistic population growth model — 264
15.2.3 Logistic map — 266
15.3 The Lotka–Volterra or predator–prey system — 270
15.4 Cellular automata — 273
15.5 Random Boolean networks — 276
15.6 Case studies of dynamical system models with complex attractors — 283
15.6.1 The Lorenz attractor — 283
15.6.2 Clifford attractor — 285
15.6.3 Ikeda attractor — 286
15.6.4 The Peter de Jong attractor — 286
17.7 Discrete and continuous distributions — 333
17.7.1 Uniform distribution — 334
17.8 Expectation values and moments — 334
17.8.1 Expectation values — 334
17.8.2 Variance — 335
17.8.3 Moments — 336
17.8.4 Covariance and correlation — 336
17.9 Bivariate distributions — 337
17.10 Multivariate distributions — 338
17.11 Important discrete distributions — 339
17.11.1 Bernoulli distribution — 339
17.11.2 Binomial distribution — 340
17.11.3 Geometric distribution — 342
17.11.4 Negative binomial distribution — 343
17.11.5 Poisson distribution — 343
17.12 Important continuous distributions — 344
17.12.1 Exponential distribution — 345
17.12.2 Beta distribution — 345
17.12.3 Gamma distribution — 346
17.12.4 Normal distribution — 347
17.12.5 Chi-square distribution — 349
17.12.6 Student’s t-distribution — 350
17.12.7 Log-normal distribution — 351
17.12.8 Weibull distribution — 352
17.13 Bayes’ theorem — 353
17.14 Information theory — 356
17.14.1 Entropy — 356
17.14.2 Kullback–Leibler divergence — 358
17.14.3 Mutual information — 359
17.15 Law of large numbers — 361
17.16 Central limit theorem — 364
17.17 Concentration inequalities — 364
17.17.1 Hoeffding’s inequality — 365
17.17.2 Cauchy–Schwartz inequality — 365
17.17.3 Chernoff bounds — 366
17.18 Further reading — 366
17.19 Summary — 367
17.20 Exercises — 367

18 Optimization — 369
18.1 Introduction — 369
18.2 Formulation of an optimization problem — 370