Contents

Foreword ................................................................. vii
List of Figures ...................................................... xxvii
List of Tables .......................................................... xxxi
Mathematical Notations ............................................... xxxiii

Part I Preliminaries

1 Introducing R ......................................................... 3
  1.1 Presentation of the Software ................................. 3
     1.1.1 Origins .............................................. 3
     1.1.2 Why Use R? .......................................... 3
  1.2 R and Statistics ............................................... 5
  1.3 R and Plots ................................................... 5
  1.4 The R Graphical User Interface .............................. 7
  1.5 First Steps in R .............................................. 7
     1.5.1 Using R Commander .................................. 7
        1.5.1.1 Launching R Commander ..................... 8
        1.5.1.2 Handling Data with R Commander ........... 8
        1.5.1.3 A Few Statistical Tasks with R Commander .. 13
        1.5.1.4 Adding Functionalities to the R Commander Interface ................................ 18
     1.5.2 Using R with the Console ............................... 19
        1.5.2.1 The Strength of R Shown on an Example ...... 19
        1.5.2.2 A Brief Introduction of R Syntax Through Some Instructions to Type ............. 23

2 A Few Data Sets and Research Questions .......................... 29
  2.1 Body Mass Index of Children ................................. 29
  2.2 Weight at Birth ............................................... 30
2.3 Intima–Media Thickness ................................................. 31
2.4 Diet of Elderly People................................................... 32
2.5 Study Case of Myocardial Infarction .............................. 33
2.6 Summary Table of Use of Data Sets ............................... 33

Part II The Bases of R

3 Basic Concepts and Data Organisation ............................... 37
3.1 Your First Session .................................................... 37
  3.1.1 R Is a Calculator .................................................. 38
  3.1.2 Displaying Results and Variable Redirecting ................. 39
  3.1.3 Work Strategy ..................................................... 40
  3.1.4 Using Functions ................................................... 43
3.2 Data in R ............................................................... 46
  3.2.1 Data Nature (or Type, or Mode) ............................... 46
    3.2.1.1 Numeric Type (numeric) ................................. 46
    3.2.1.2 Complex Type (complex) ............................... 47
    3.2.1.3 Boolean or Logical Type (logical) ..................... 48
    3.2.1.4 Missing Data (NA) ........................................ 48
    3.2.1.5 Character String Type (character) ...................... 49
    3.2.1.6 Raw Data (raw) ........................................... 50
  Summary ......................................................................... 50
  3.2.2 Data Structures .................................................... 50
    3.2.2.1 Vectors (vector) ............................................. 51
    3.2.2.2 Matrices (matrix) and Arrays (array) ................. 52
    3.2.2.3 Lists (list) ................................................... 53
    3.2.2.4 The Individual×Variable Table (data.frame) .......... 55
    3.2.2.5 Factors (factor), Ordinal Variables (ordered) ........ 56
  3.2.2.6 Dates ............................................................ 57
  3.2.2.7 Time Series ..................................................... 57
  Summary ......................................................................... 58
Memorandum ...................................................................... 59
Exercises .......................................................................... 59
Worksheet ......................................................................... 60

4 Importing, Exporting and Producing Data ........................... 63
4.1 Importing Data ........................................................... 63
  4.1.1 Importing Data from an ASCII Text File ..................... 63
    4.1.1.1 Reading Data with read.table() ......................... 64
    4.1.1.2 Reading Data with read.ftable() ....................... 67
    4.1.1.3 Reading Data with the Function scan() ............... 68
4.1.2 Importing Data from Excel or the Open Office Spreadsheet ..................................... 69
  4.1.2.1 Copy-Pasting ................................ 69
  4.1.2.2 Using an Intermediary ASCII File ............... 70
  4.1.2.3 Using Specialized Packages ..................... 70
4.1.3 Importing Data from SPSS, Minitab, SAS or Matlab ............................................. 70
4.1.4 Large Data Files .......................................................... 71
4.2 Exporting Data ................................................................. 72
  4.2.1 Exporting Data to an ASCII Text File ............... 72
  4.2.2 Exporting Data to Excel or OpenOffice Calc ............ 72
4.3 Creating Data ................................................................. 73
  4.3.1 Entering Toy Data ................................................ 73
  4.3.2 Generating Pseudo-Random Numbers ................... 74
  4.3.3 Entering Data from a Hard Copy ....................... 74
4.4 Reading/Writing in Databases ............................................. 76
  4.4.1 Creating a Database and a Table ....................... 76
  4.4.2 Creating a Data Source Compatible with MySQL ......... 76
  4.4.3 Writing in a Table ................................................ 78
  4.4.4 Reading a Table ................................................ 79

Memorandum ............................................................................. 80
Exercises ................................................................................ 80
Worksheet ................................................................................ 81

5 Data Manipulation, Functions .................................................. 85
5.1 Operations on Vectors, Matrices and Lists .................................. 85
  5.1.1 Vector Arithmetic .................................................... 85
  5.1.2 Recycling ................................................................. 86
  5.1.3 Basic Functions ......................................................... 87
  5.1.4 Operations on Matrices and DataFrames ....................... 88
    5.1.4.1 Information on Architecture .......................... 88
    5.1.4.2 Merging Tables .................................................. 89
    5.1.4.3 The Function apply() ....................................... 93
    5.1.4.4 The Function sweep() ....................................... 94
    5.1.4.5 The Function stack() ......................................... 94
    5.1.4.6 The Function aggregate() ................................. 95
    5.1.4.7 The Function transform() ................................. 95
  5.1.5 Operations on Lists ...................................................... 96
5.2 Logical and Relational Operations ............................................. 97
5.3 Operations on Sets .............................................................. 98
5.4 Extracting and Inserting Elements ........................................... 99
  5.4.1 Extracting from/Inserting into Vectors ...................... 100
  5.4.2 Extracting from/Inserting into Matrices .................. 102
  5.4.3 Extracting from/Inserting into Arrays .................... 106
  5.4.4 Extracting from/Inserting into Lists .................... 106
5.5 Manipulating Character Strings ............................................. 108
5.6 Manipulating Dates and Time Units ........................................... 111
  5.6.1 Displaying the Current Date ............................................... 111
  5.6.2 Extracting Dates ............................................................. 112
  5.6.3 Operations on Dates .......................................................... 113

5.7 Control Flow ............................................................................. 115
  5.7.1 Conditional Instructions ......................................................... 116
  5.7.2 Loop Instructions .................................................................. 118

5.8 Creating Functions ................................................................. 120

5.9 Fixed-Point and Floating Point Number Representation ............ 127
  5.9.1 Representing a Number in a Base ............................................. 127
  5.9.2 Floating Point Representations ............................................... 128
    5.9.2.1 Definitions ..................................................................... 128
    5.9.2.2 Limitations of This Representation due to the Significand ...... 129
    5.9.2.3 Avoiding Some Numerical Pitfalls ...................................... 130
    5.9.2.4 Limitations of This Representation due to the Exponent ...... 132

Memorandum .................................................................................. 134
Exercises ......................................................................................... 134
Worksheet ....................................................................................... 136

6 R and Its Documentation .............................................................. 141
  6.1 Integrated Help ......................................................................... 141
    6.1.1 The Command help() .......................................................... 141
    6.1.2 Some Complementary Commands ......................................... 143

6.2 Help on the Web ......................................................................... 145
    6.2.1 Search Engines .................................................................. 145
    6.2.2 Message Boards ................................................................. 146
    6.2.3 Mailing Lists ...................................................................... 146
    6.2.4 Internet Relay Chat (IRC) ..................................................... 147
    6.2.5 Wiki .................................................................................... 147

6.3 Literature About R ...................................................................... 147
    6.3.1 Online ................................................................................ 147
    6.3.2 Printed Material ................................................................. 148

Memorandum .................................................................................. 149
Exercises ......................................................................................... 149
Worksheet ....................................................................................... 149

7 Drawing Curves and Plots ............................................................ 151
  7.1 Graphics Windows ..................................................................... 151
    7.1.1 Basic Graphics Windows, Manipulation and Saving .............. 151
    7.1.2 Splitting the Graphics Window: layout() ............................. 153

7.2 Low-Level Drawing Functions .................................................. 156
    7.2.1 The Functions plot() and points() ...................................... 156
    7.2.2 The Functions segments(), lines() and abline() ...................... 158
8.4 † Going Further in R Programming ................................................. 216
  8.4.1 R Attributes ...................................................................... 216
     8.4.1.1 Attribute class ..................................................... 218
     8.4.1.2 Attribute dim ....................................................... 218
     8.4.1.3 Attributes names and dimnames ....................... 221
  8.4.2 Other R Objects ............................................................ 224
     8.4.2.1 R Expressions ..................................................... 224
     8.4.2.2 R Formulae .......................................................... 226
     8.4.2.3 The R Environment .............................................. 228
  8.5 † Interfacing R and C/C++ or Fortran ...................................... 230
     8.5.1 Creating and Running a C/C++ or Fortran Function ........ 231
     8.5.2 Calling C/C++ (or Fortran) from R ............................. 237
     8.5.3 Calling External C/C++ or Fortran Libraries ............... 242
        8.5.3.1 The R API ....................................................... 243
        8.5.3.2 The newmat Library .......................................... 246
        8.5.3.3 The BLAS and LAPACK Packages ...................... 248
        8.5.3.4 Mixing C/C++ and Fortran Packages ................. 250
     8.5.4 Calling R Code from a C/C++ Program Called by R .... 252
     8.5.5 Calling R Code from Fortran ........................................ 255
     8.5.6 Some Useful Functions ............................................. 255
  8.6 † Debugging Functions .......................................................... 255
     8.6.1 Debugging Functions in Pure R .................................... 255
     8.6.2 Error in R Code ........................................................ 257
     8.6.3 Error in the C/C++ or Fortran Code ......................... 258
     8.6.4 Debugging with GDB ................................................ 259
        8.6.4.1 Debugging with Emacs ...................................... 262
        8.6.4.2 Debugging with DDD ......................................... 264
        8.6.4.3 Debugging with Insight .................................... 265
        8.6.4.4 Detecting Memory Leaks ................................. 270
  8.7 Parallel Computing and Computation on Graphical Cards ............. 273
     8.7.1 Parallel Computing ................................................... 273
     8.7.2 Computation on Graphical Cards ............................... 274
  Memorandum ............................................................................ 276
  Exercises .............................................................................. 276
  Worksheet ........................................................................... 278

9 Managing Sessions ................................................................. 283
  9.1 R Commands, Objects and Storage ....................................... 283
  9.2 Workspace: .RData Files .................................................... 285
  9.3 Command History: .Rhistory Files ..................................... 287
  9.4 Saving Plots .................................................................... 288
  9.5 Managing Packages .......................................................... 290
  9.6 Managing Access Paths to R Objects .................................. 290
  9.7 † Other Useful Commands ................................................. 292
Contents

9.8 Problems in Memory Management ................................................................. 293
  9.8.1 Organization of RAM ........................................................................... 293
  9.8.2 Accessing the Memory ......................................................................... 294
    9.8.2.1 Problems Caused by Memory Management of Integers ............... 295
    9.8.2.2 Successive Allocation of Memory .............................................. 296
  9.8.3 Object Size in R ................................................................................. 298
  9.8.4 Total Memory used by R ...................................................................... 299
  9.8.5 A Few Recommendations .................................................................... 301

9.9 Using R in BATCH Mode ........................................................................... 302
9.10 Creating a Simple R Package ................................................................. 303

Memorandum .................................................................................................. 306
Exercises ........................................................................................................ 306
Worksheet ....................................................................................................... 307

Part III Elementary Mathematics and Statistics

10 Basic Mathematics: Matrix Operations, Integration, Optimization ............... 313
  10.1 Basic Mathematical Functions ............................................................... 313
  10.2 Matrix Operations ................................................................................. 315
    10.2.1 Basic Matrix Operations .................................................................. 316
    10.2.2 Outer Product ................................................................................. 318
    10.2.3 Kronecker Product .......................................................................... 319
    10.2.4 Triangular Matrices ........................................................................ 319
    10.2.5 Operators vec and Half vec ............................................................ 320
    10.2.6 Determinant, Trace and Condition Number .................................... 320
    10.2.7 Scaling and Centring Data ............................................................... 321
    10.2.8 Eigenvalues and Eigenvectors ......................................................... 321
    10.2.9 Square Root of a Hermitian Positive-Definite Matrix ....................... 322
    10.2.10 Singular Value Decomposition ...................................................... 323
    10.2.11 Cholesky Decomposition ............................................................... 323
    10.2.12 QR Decomposition ........................................................................ 324
  10.3 Numerical Integration ............................................................................ 325
  10.4 Differentiation ......................................................................................... 326
    10.4.1 Symbolic Differentiation .................................................................. 326
    10.4.2 Numerical Differentiation ............................................................... 327
  10.5 Optimization ........................................................................................... 327
    10.5.1 Optimization Functions ................................................................... 327
    10.5.2 Roots of a Function .......................................................................... 331

Memorandum .................................................................................................. 333
Exercises ........................................................................................................ 333
Worksheet ....................................................................................................... 334
11 Descriptive Statistics ............................................. 339

11.1 Introduction .................................................. 339

11.2 Structuring Variables According to Type .................... 340
  11.2.1 Structuring Qualitative Variables .................... 341
  11.2.2 Structuring Ordinal Variables ....................... 342
  11.2.3 Structuring Discrete Quantitative Data ............. 342
  11.2.4 Structuring Continuous Quantitative Variables ...... 343

11.3 Data Tables .................................................. 343
  11.3.1 Individual Data Tables ................................ 343
  11.3.2 Tables of Counts and Frequency Tables .......... 343
  11.3.3 Tables of Grouped Data ............................ 344
  11.3.4 Cross Tabulation ..................................... 344
    11.3.4.1 Contingency Tables .......................... 344
    11.3.4.2 Joint Distribution .......................... 345
    11.3.4.3 Marginal Distributions ..................... 346
    11.3.4.4 Conditional Distributions .................. 346

11.4 Numerical Summaries ........................................ 347
  11.4.1 Summaries of the Location of a Distribution ......... 348
    11.4.1.1 Modes .................................... 348
    11.4.1.2 Median .................................... 348
    11.4.1.3 Mean ..................................... 350
    11.4.1.4 Quantiles .................................. 350
  11.4.2 Summaries of the Dispersion of a Distribution ...... 350
  11.4.3 Summaries of the Shape of a Distribution .......... 351

11.5 Measures of Association ................................... 352
  11.5.1 Measures of Association Between Two Qualitative Variables ............................................. 352
    11.5.1.1 Pearson’s $\chi^2$ Statistic .................. 352
    11.5.1.2 $\phi^2$, Cramér’s $V$ and Pearson’s Contingency Coefficient .......................... 353
  11.5.2 Measures of Association Between Ordinal Variables (or Ranks) .................................................. 354
    11.5.2.1 Kendall’s $\tau$ and $\tau_b$ ................. 354
    11.5.2.2 Spearman’s Rank Correlation Coefficient $\rho$ ............................................. 355
  11.5.3 Measures of Association Between Two Quantitative Variables ............................................. 355
    11.5.3.1 Covariance and Pearson’s Correlation Coefficient ........................................... 355
  11.5.4 Measures of Association Between a Quantitative Variable and a Qualitative Variable ................. 356
    11.5.4.1 Correlation Ratio $\eta^2_{Y|X}$ .................. 356

11.6 Graphical Representations ................................... 357
  11.6.1 Plotting Qualitative Variables ....................... 357
    11.6.1.1 Cross Chart .................................. 357
    11.6.1.2 Bar Charts ................................... 359
11.6.1.3 Pareto Chart ........................................ 360
11.6.1.4 Stacked Bar Chart .................................. 361
11.6.1.5 Pie Chart ........................................... 361
11.6.2 Plotting Ordinal Variables .......................... 362
11.6.2.1 Bar Chart with Cumulative Frequencies Line .......................... 362
11.6.3 Plotting Discrete Quantitative Variables ...................... 363
11.6.3.1 Cross Chart ....................................... 363
11.6.3.2 Bar Chart .......................................... 363
11.6.3.3 Plotting the Empirical Distribution Function .................. 363
11.6.3.4 Stemplot ........................................... 365
11.6.3.5 Boxplot ............................................ 365
11.6.4 Plotting Continuous Quantitative Variables ............... 367
11.6.4.1 Empirical Distribution Function ..................... 367
11.6.4.2 Stemplot ........................................... 367
11.6.4.3 Boxplots ............................................ 368
11.6.4.4 Density Histogram with Identical or Different Class Ranges .................. 368
11.6.4.5 Frequency Polygon .................................. 369
11.6.4.6 Cumulative Frequency Polygon ....................... 370
11.6.5 Graphical Representations in a Bivariate Setting .......... 371
11.6.5.1 Two-Way Plots for Two Qualitative Variables ............. 371
11.6.5.2 Two-way Plots for Two Quantitative Variables ............... 374
11.6.5.3 Two-Way Plots for One Qualitative and One Quantitative Variable .................. 375
Memorandum ................................................ 377
Exercises .................................................. 377
Worksheet .................................................. 378

12 A Better Understanding of Random Variables, Distributions and Simulations Using R Specificities .................. 381
12.1 Notions on Random Number Generation ...................... 381
12.2 The Notion of Random Variables .............................. 383
12.2.1 Realizations of a Random Variable and Functioning Law .................. 383
12.2.2 I.i.d. Random Variables .................................. 384
12.2.3 Characterizing the Distribution of a Random Variable ............. 385
12.2.3.1 Density Function, Distribution Function and Quantile Function .................. 387
12.2.4 Parameters of the Distribution of a Random Distribution .......... 390
12.3 Law of Large Numbers and Central Limit Theorem ............. 392
12.3.1 Law of Large Numbers .................................. 392
12.3.2 Central Limit Theorem .................................. 393
12.4 Inferential Statistics ...................................... 394
  12.4.1 Point Estimate of Parameters ............................ 394
  12.4.2 Empirical Cumulative Distribution Function ............. 396
  12.4.3 Maximum Likelihood Estimation .......................... 397
  12.4.4 Sampling Variation and Properties of an Estimator .... 399
12.5 A Few Techniques to Draw from a Distribution .............. 401
  12.5.1 Simulating from Another Distribution ...................... 401
  12.5.2 Inverse Transform Method .............................. 402
  12.5.3 Rejection Sampling .................................. 402
  12.5.4 Simulation of Discrete Random Variables .............. 403
12.6 Bootstrap .............................................. 404
12.7 Standard and Less Standard Distributions ................... 405
  12.7.1 Standard Distributions ................................ 405
  12.7.2 Less Standard Distributions ............................ 408
12.8 Modelling a Phenomenon .................................... 410
Memorandum ............................................... 413
Exercises ................................................... 413
Worksheet .................................................. 413

13 Confidence Intervals and Hypothesis Testing ................. 417
  13.1 Notations .............................................. 417
  13.2 Confidence Intervals ..................................... 418
    13.2.1 Confidence Intervals for the Mean .................. 418
    13.2.2 Confidence Intervals for a Proportion $p$ ........... 419
    13.2.3 Confidence Intervals for a Variance ............... 421
    13.2.4 Confidence Intervals for a Median .................. 422
    13.2.5 Confidence Intervals for a Correlation Coefficient 423
    13.2.6 Summary Table for Confidence Intervals ............. 424
  13.3 Standard Hypothesis Testing .............................. 424
    13.3.1 Parametric Tests .................................. 426
      13.3.1.1 Tests of the Mean ................................ 426
      13.3.1.2 Tests of Variance ................................ 429
      13.3.1.3 Tests of Proportion ............................. 431
      13.3.1.4 Tests of Correlation ............................. 433
    13.3.2 Independence Tests ................................ 435
      13.3.2.1 $\chi^2$ Test for Independence .................... 435
      13.3.2.2 Yates’ $\chi^2$ Test ............................ 437
      13.3.2.3 Fisher’s Exact Test ............................. 438
    13.3.3 Non-parametric Tests ............................... 439
      13.3.3.1 Goodness-of-Fit Tests ........................... 439
      13.3.3.2 Tests of Position ............................... 442
    13.3.4 Memorandum of Standard Tests ....................... 447
  13.4 Other Tests ........................................... 447
Memorandum ............................................... 449
Exercises ................................................... 449
Worksheet .................................................. 449
14 Simple and Multiple Linear Regression ........................................ 455
  14.1 Introduction .................................................. 455
  14.2 Simple Linear Regression ....................................... 456
    14.2.1 Aim and Model ...................................... 456
    14.2.2 Fitting Data ......................................... 457
    14.2.3 Confidence and Prediction Intervals for a New Value ... 461
    14.2.4 Analysis of Residuals .................................. 463
    14.2.5 Student’s Tests for Means and Linear Model ............ 466
    14.2.6 Summary ............................................... 468
  14.3 Multiple Linear Regression ..................................... 468
    14.3.1 Aim and Model ...................................... 468
    14.3.2 Fitting Data ......................................... 469
    14.3.3 Confidence and Prediction Intervals for a New Value ... 473
    14.3.4 Testing a Linear Sub-hypothesis: Partial Fisher Test . . 473
    14.3.5 Qualitative Variables with More Than Two Modalities ........................................ 474
    14.3.6 Interaction Between Variables .......................... 478
    14.3.7 Issues with Collinearity .............................. 481
    14.3.8 Variable Selection .................................... 482
    14.3.9 Analysis of Residuals .................................. 490
    14.3.10 Polynomial Regression .............................. 496
    14.3.11 Summary ............................................... 496
Memorandum ............................................................................ 497
Exercises ............................................................................... 497
Worksheet .............................................................................. 498

15 Elementary Analysis of Variance .......................................... 503
  15.1 Analysis of Variance with One Factor ............................. 503
    15.1.1 Aims, Data and Model ................................... 503
    15.1.2 Example and Graphical Inspection ....................... 504
    15.1.3 ANOVA Table and Parameter Estimation ................. 505
    15.1.4 Validation of Assumptions ............................... 509
    15.1.5 Multiple Comparisons and Contrasts ..................... 510
    15.1.6 Summary .................................................. 512
  15.2 Analysis of Variance with Two Factors ............................ 513
    15.2.1 Aims, Data and Model ................................... 513
    15.2.2 Example and Graphical Inspection ....................... 514
    15.2.3 ANOVA Table, Tests and Parameter Estimation ........ 516
    15.2.4 Validating Assumptions .................................. 519
    15.2.5 Contrasts .................................................. 519
    15.2.6 Summary .................................................. 521
  15.3 Repeated Measures Analysis of Variance .......................... 521
    15.3.1 One-Way Repeated Measures ANOVA ..................... 522
    15.3.2 Two-Factor Model with Repeated Measures for Both Factors ........................................ 523
15.3.3 Two-Factor Model with Repeated Measures for One Factor .......................................... 525
Memorandum ............................................... 527
Exercises ................................................... 527
Worksheet .................................................. 527

Appendix: Installing R and R Packages ............................ 531
A.1 Installing R Under Microsoft Windows ..................... 531
A.2 Installing Additional Packages ............................. 532
    A.2.1 Installing from a File on Your Disk ................. 532
    A.2.2 Installing Directly from the Internet ................. 533
    A.2.3 Installing from the Command Line ................. 535
    A.2.4 Installing Packages Under Linux ................... 535
A.3 Loading Installed Packages ............................... 536

References ...................................................... 539

General Index .................................................. 541

Index of R Commands and Symbols .................................. 549

Index of Authors ................................................ 561

List of R Packages Mentioned in the Book .......................... 563

Solutions to Exercises ............................................ 565

Solutions to Worksheet ........................................... 577
The R Software
Fundamentals of Programming and Statistical Analysis
Lafaye de Micheaux, P.; Drouilhet, R.; Liquet, B.
2013, XXXVII, 628 p. 50 illus., 41 illus. in color.,
Hardcover
ISBN: 978-1-4614-9019-7