Statistics and Computing

Peter Dalgaard

Introductory Statistics with R

Second Edition

Springer

Contents

Preface **vii**

1 Basics **1**

1.1 First steps 1

1.1.1 An overgrown calculator 3

1.1.2 Assignments 3

1.1.3 Vectorized arithmetic 4

1.1.4 Standard procedures 6

1.1.5 Graphics 7

1.2 R language essentials 9

1.2.1 Expressions and objects 9

1.2.2 Functions and arguments 11

1.2.3 Vectors 12

1.2.4 Quoting and escape sequences 13

1.2.5 Missing values 14

1.2.6 Functions that create vectors 14

1.2.7 Matrices and arrays 16

1.2.8 Factors 18

1.2.9 Lists 19

1.2.10 Data frames 20

1.2.11 Indexing 21

1.2.12 Conditional selection 22

1.2.13 Indexing of data frames 23

1.2.14 Grouped data and data frames 25

1.2.15 Implicit loops 26
1.2.16 Sorting 27
1.3 Exercises 28

2 The R environment 31
2.1 Session management 31
2.1.1 The workspace 31
2.1.2 Textual output 32
2.1.3 Scripting 33
2.1.4 Getting help 34
2.1.5 Packages 35
2.1.6 Built-in data 35
2.1.7 attach and detach 36
2.1.8 subset, transform, and within 37
2.2 The graphics subsystem 39
2.2.1 Plot layout 39
2.2.2 Building a plot from pieces 40
2.2.3 Using par 42
2.2.4 Combining plots 42
2.3 R programming 44
2.3.1 Flow control 44
2.3.2 Classes and generic functions 46
2.4 Data entry 46
2.4.1 Reading from a text file 47
2.4.2 Further details on read.table 50
2.4.3 The data editor 51
2.4.4 Interfacing to other programs 52
2.5 Exercises 53

3 Probability and distributions 55
3.1 Random sampling 55
3.2 Probability calculations and combinatorics 56
3.3 Discrete distributions 57
3.4 Continuous distributions 58
3.5 The built-in distributions in R 59
3.5.1 Densities 59
3.5.2 Cumulative distribution functions 62
3.5.3 Quantiles 63
3.5.4 Random numbers 64
3.6 Exercises 65

4 Descriptive statistics and graphics 67
4.1 Summary statistics for a single group 67
4.2 Graphical display of distributions 71
4.2.1 Histograms 71

4.2.2 Empirical cumulative distribution 73
4.2.3 Q–Q plots 74
4.2.4 Boxplots 75
4.3 Summary statistics by groups 75
4.4 Graphics for grouped data 79
4.4.1 Histograms 79
4.4.2 Parallel boxplots 80
4.4.3 Stripcharts 81
4.5 Tables 83
4.5.1 Generating tables 83
4.5.2 Marginal tables and relative frequency 87
4.6 Graphical display of tables 89
4.6.1 Barplots 89
4.6.2 Dotcharts 91
4.6.3 Piecharts 92
4.7 Exercises 93

5 One- and two-sample tests 95
5.1 One-sample *t* test 95
5.2 Wilcoxon signed-rank test 99
5.3 Two-sample *t* test 100
5.4 Comparison of variances 103
5.5 Two-sample Wilcoxon test 103
5.6 The paired *t* test 104
5.7 The matched-pairs Wilcoxon test 106
5.8 Exercises 107

6 Regression and correlation 109
6.1 Simple linear regression 109
6.2 Residuals and fitted values 113
6.3 Prediction and confidence bands 117
6.4 Correlation 120
6.4.1 Pearson correlation 121
6.4.2 Spearman's ρ 123
6.4.3 Kendall's τ 124
6.5 Exercises 124

7 Analysis of variance and the Kruskal–Wallis test 127
7.1 One-way analysis of variance 127
7.1.1 Pairwise comparisons and multiple testing . . . 131
7.1.2 Relaxing the variance assumption 133
7.1.3 Graphical presentation 134
7.1.4 Bartlett's test 136
7.2 Kruskal–Wallis test 136
7.3 Two-way analysis of variance 137

7.3.1 Graphics for repeated measurements 140
7.4 The Friedman test 141
7.5 The ANOVA table in regression analysis 141
7.6 Exercises 143

8 Tabular data 145
8.1 Single proportions 145
8.2 Two independent proportions 147
8.3 *k* proportions, test for trend 149
8.4 $r \times c$ tables 151
8.5 Exercises 153

9 Power and the computation of sample size 155
9.1 The principles of power calculations 155
9.1.1 Power of one-sample and paired *t* tests 156
9.1.2 Power of two-sample *t* test 158
9.1.3 Approximate methods 158
9.1.4 Power of comparisons of proportions 159
9.2 Two-sample problems 159
9.3 One-sample problems and paired tests 161
9.4 Comparison of proportions 161
9.5 Exercises 162

10 Advanced data handling 163
10.1 Recoding variables 163
10.1.1 The `cut` function 163
10.1.2 Manipulating factor levels 165
10.1.3 Working with dates 166
10.1.4 Recoding multiple variables 169
10.2 Conditional calculations 170
10.3 Combining and restructuring data frames 171
10.3.1 Appending frames 172
10.3.2 Merging data frames 173
10.3.3 Reshaping data frames 175
10.4 Per-group and per-case procedures 178
10.5 Time splitting 179
10.6 Exercises 183

11 Multiple regression 185
11.1 Plotting multivariate data 185
11.2 Model specification and output 187
11.3 Model search 190
11.4 Exercises 193

12 Linear models 195
12.1 Polynomial regression 196
12.2 Regression through the origin 198
12.3 Design matrices and dummy variables 200
12.4 Linearity over groups 202
12.5 Interactions 206
12.6 Two-way ANOVA with replication 207
12.7 Analysis of covariance 208
12.7.1 Graphical description 209
12.7.2 Comparison of regression lines 212
12.8 Diagnostics 218
12.9 Exercises 224

13 Logistic regression 227
13.1 Generalized linear models 228
13.2 Logistic regression on tabular data 229
13.2.1 The analysis of deviance table 234
13.2.2 Connection to test for trend 235
13.3 Likelihood profiling 237
13.4 Presentation as odds-ratio estimates 239
13.5 Logistic regression using raw data 239
13.6 Prediction 241
13.7 Model checking 242
13.8 Exercises 247

14 Survival analysis 249
14.1 Essential concepts 249
14.2 Survival objects 250
14.3 Kaplan–Meier estimates 251
14.4 The log-rank test 254
14.5 The Cox proportional hazards model 256
14.6 Exercises 258

15 Rates and Poisson regression 259
15.1 Basic ideas 259
15.1.1 The Poisson distribution 260
15.1.2 Survival analysis with constant hazard 260
15.2 Fitting Poisson models 262
15.3 Computing rates 266
15.4 Models with piecewise constant intensities 270
15.5 Exercises 274

16 Nonlinear curve fitting 275
16.1 Basic usage 276
16.2 Finding starting values 278

16.3 Self-starting models 284
16.4 Profiling 285
16.5 Finer control of the fitting algorithm 287
16.6 Exercises 288

A Obtaining and installing R and the `ISwR` package 289

B Data sets in the `ISwR` package 293

C Compendium 325

D Answers to exercises 337

Bibliography 355

Index 357