

Contents

Preface	xv
1 Getting started	1
1.1 What is data?	1
1.2 Getting started with R	3
Installing R	3
Installing RSTUDIO	4
R's command line	5
Variables	7
Functions	8
The workspace	12
External packages	15
Data sets	16
Problems	18
2 Univariate data	20
2.1 Data vectors	22
Structured data	28
Indexing	29
Data types	33
Numeric data types	33
Categorical data types	34
Date and time types	39
Logical data	41
Problems	45
2.2 Functions	48
Problems	50
2.3 Numeric summaries	50
Center	51
The sample mean	51
The sample median	55
Measures of position	56
Other measures of center	59

Spread	59
The variance and standard deviation	60
The IQR	65
Shape	66
Viewing the shape of a data set	70
Problems	81
2.4 Categorical data	85
Problems	87
3 Bivariate data	88
3.1 Independent samples	88
Problems	93
3.2 Data manipulation basics	94
Lists	94
Data frames	96
Model formulas	97
Problems	101
3.3 Paired data	102
Correlation	105
Trends	115
Transformations	120
Alternative trend lines	123
Problems	128
3.4 Bivariate categorical data	132
Tables	132
Two-way tables from summarized data	132
Two-way tables from unsummarized data	134
Marginal distributions of two-way tables	135
Conditional distributions of two-way tables	136
The xtabs function	137
Graphical summaries of two-way contingency tables	140
Mosaic plots	141
Measures of association for categorical data	143
Problems	149
4 Multivariate data	150
4.1 Data structures in R	150
Problems	154
4.2 Working with data frames	155
Problems	166
4.3 Applying a function over a collection	167
Map	168
Filter	177
Reduce	177
Problems	179
4.4 Using external data	181

Spreadsheet data	181
Web-based data sets	182
5 Multivariate graphics	189
5.1 Base graphics	189
Problems	196
5.2 Lattice graphics	197
Problems	200
5.3 The ggplot2 package	200
Geoms	201
Grouping	203
Statistical transformations	204
Faceting	207
Problems	210
6 Populations	211
6.1 Populations	211
Discrete random variables	213
Using <code>sample</code> to generate random values	214
The mean and standard deviation	215
Continuous random variables	216
The p.d.f. and c.d.f.	218
The mean and standard deviation	218
Quantiles	218
Sampling from a population	219
Random samples generated by <code>sample</code>	219
Sampling distributions	220
Problems	221
6.2 Families of distributions	222
The <code>d</code> , <code>p</code> , <code>q</code> , and <code>r</code> functions	222
Binomial, normal, and some other named distributions	224
Bernoulli random variables	224
Binomial random variables	225
Normal random variables	227
Popular distributions to describe populations	231
Uniform distribution	231
Exponential distribution	232
Lognormal distribution	233
Sampling distributions	233
Problems	234
6.3 The central limit theorem	236
Normal parent population	237
Nonnormal parent population	238
Problems	240

7 Statistical inference	242
7.1 Simulation	244
Repeating a simulation easily	244
Problems	252
7.2 Significance tests	252
7.3 Estimation, confidence intervals	255
The basic bootstrap	258
7.4 Bayesian analysis	259
8 Confidence intervals	262
8.1 Confidence intervals for a population proportion, p	264
Problems	269
8.2 Confidence intervals for the population mean	271
One-sided confidence intervals	274
Problems	276
8.3 Other confidence intervals	278
Confidence interval for σ^2	278
Problems	280
8.4 Confidence intervals for differences	281
Difference of proportions	282
Difference of means	283
Matched samples	286
Problems	287
8.5 Confidence intervals for the median	288
Confidence intervals based on the binomial distribution	288
Confidence intervals based on signed-rank statistic	289
Confidence intervals based on the rank-sum statistic	290
Problems	292
9 Significance tests	294
9.1 Significance test for a population proportion	299
Using prop. test to compute p -values	301
Problems	302
9.2 Significance test for the mean (t -tests)	304
Power	307
Problems	309
9.3 Significance tests and confidence intervals	310
9.4 Significance tests for the median	312
The sign test	312
The signed-rank test	313
Problems	315
9.5 Two-sample tests of proportion	316
Problems	319
9.6 Two-sample tests of center	321
Two-sample tests of center with normal populations	322
Matched samples	325

The Wilcoxon rank-sum test for equality of center	328
Problems	331
10 Goodness of fit	334
10.1 The chi-squared goodness-of-fit test	334
The multinomial distribution	334
Pearson's χ^2 -statistic	336
Partially specified null hypotheses	339
Problems	341
10.2 The chi-squared test of independence	344
The chi-squared test of homogeneity	348
Problems	350
10.3 Goodness-of-fit tests for continuous distributions	352
Kolmogorov-Smirnov test	352
The Shapiro-Wilk test for normality	357
Finding parameter values using fitdistr	359
Problems	362
11 Linear regression	364
11.1 The simple linear regression model	364
Estimating the parameters in simple linear regression	365
Using lm to find the estimates	366
Extractor functions for lm	367
Problems	368
11.2 Statistical inference for simple linear regression	369
Statistical inferences	370
Marginal t -tests	370
The F -test	371
R^2 —the coefficient of determination	373
Using lm to find values for a regression model	374
Confidence intervals	374
Standard error	374
Significance tests	376
Finding $\hat{\sigma}^2$, R^2	376
F -test for $\beta_1 = 0$	377
Predicting the response with predict	377
Testing the model assumptions	378
Assessing the linear model for the mean	379
Assessing the residuals	380
Influential points	381
Prediction intervals	382
Confidence intervals for $\mu_{y x}$	385
Problems	386
11.3 Multiple linear regression	390
Types of models	390
Fitting the multiple regression model using lm	392

Using update with model formulas	394
Interpreting the regression parameters	395
Statistical inferences	396
Model selection	397
Partial F -test	398
The Akaike information criterion	400
Problems	402
12 Analysis of variance	404
12.1 One-way ANOVA	404
Using R's model formulas to specify ANOVA models	408
Using oneway.test to perform ANOVA	408
Using aov for ANOVA	409
The nonparametric Kruskal–Wallis test	411
Problems	414
12.2 Using lm for ANOVA	416
Treatment coding for analysis of variance	418
Comparing multiple differences	421
Problems	424
12.3 ANCOVA	425
Problems	428
12.4 Two-way ANOVA	429
Interaction plots	430
Fitting a two-way ANOVA	431
Blocking variables	435
Problems	437
13 Extensions of the linear model	440
13.1 Logistic regression	440
Generalized linear models	443
Fitting the model using glm	443
13.2 Nonlinear models	448
Fitting nonlinear models with nls	449
Problems	455
A Programming	458
A.1 Functions	458
Function names	458
Arguments	462
Body	467
Control flow	467
Variable scope	472
Closures	474
A.2 Generic functions	475
S3 methods	475
S4 classes and methods	479

CONTENTS

xiii

Reference classes

479

Bibliography

489

Index

494

Preface

About this book

This is a second edition of a book that introduces R alongside the introductory statistics curriculum. The first edition found its niche with individuals looking to get started with both areas outside of a classroom environment. It is the hope, that this second edition will be even more useful for that task.

The book was first published in 2004, when R was at version 2.0.0. Now, as of writing, R is past version 3.0.0 (3.1.0 and climbing). In that time so much has changed. For example:

- The number of R users has grown enormously. A recent survey ranked R the 13th most used programming language.
- The number of add-on packages for R has grown four- or five-fold to over 5,000. The depth and range of applications has grown considerably.
- The number of books including material on R has grown at least ten-fold.¹
- The internet has developed many additional R communities beyond the initial mailing list. Two key additions are the question and answer site stackoverflow.com which has nearly 50,000 questions tagged with “r” and the blog aggregator r-bloggers.com which has over 13,900 blog entries related to R.

Basically, the amount of material out there related to learning and using R is now enormous. This book doesn’t try to canvas even a sliver, rather it tries to guide the reader through the learning of the basics of R so that it is possible to take advantage of the contributions made by the R community. Though R—like other programming languages—has a reputation of having

¹For example, there are many other texts introducing R, as this one does, that can be chosen to learn from. For example, [35], [64], [13], [14], [36], [12], [56], and <http://www.camintro.org/stat/>.