

---

# Contents

<b>Preface</b>	xv
<b>Part I Data in statistics and R</b>	
<b>1 Basic R</b>	<b>3</b>
1.1 Preliminaries	4
1.1.1 An R session	4
1.1.2 Editing statements	8
1.1.3 The functions <code>help()</code> , <code>help.search()</code> and <code>example()</code>	8
1.1.4 Expressions	10
1.1.5 Comments, line continuation and <code>Esc</code>	11
1.1.6 <code>source()</code> , <code>sink()</code> and <code>history()</code>	11
1.2 Modes	13
1.3 Vectors	14
1.3.1 Creating vectors	14
1.3.2 Useful vector functions	15
1.3.3 Vector arithmetic	15
1.3.4 Character vectors	17
1.3.5 Subsets and index vectors	18
1.4 Arithmetic operators and special values	20
1.4.1 Arithmetic operators	20
1.4.2 Logical operators	21
1.4.3 Special values	22
1.5 Objects	24
1.5.1 Orientation	24
1.5.2 Object attributes	26
1.6 Programming	28
1.6.1 Execution controls	28
1.6.2 Functions	30
1.7 Packages	33

1.8	Graphics	34
1.8.1	High-level plotting functions	35
1.8.2	Low-level plotting functions	36
1.8.3	Interactive plotting functions	36
1.8.4	Dynamic plotting	36
1.9	Customizing the workspace	36
1.10	Projects	37
1.11	A note about producing figures and output	39
1.11.1	<code>openg()</code>	39
1.11.2	<code>saveg()</code>	40
1.11.3	<code>h()</code>	40
1.11.4	<code>nqd()</code>	40
1.12	Assignments	41
<b>2</b>	<b>Data in statistics and in R</b>	<b>45</b>
2.1	Types of data	45
2.1.1	Factors	45
2.1.2	Ordered factors	48
2.1.3	Numerical variables	49
2.1.4	Character variables	50
2.1.5	Dates in R	50
2.2	Objects that hold data	50
2.2.1	Arrays and matrices	51
2.2.2	Lists	52
2.2.3	Data frames	54
2.3	Data organization	55
2.3.1	Data tables	55
2.3.2	Relationships among tables	57
2.4	Data import, export and connections	58
2.4.1	Import and export	58
2.4.2	Data connections	60
2.5	Data manipulation	63
2.5.1	Flat tables and expand tables	63
2.5.2	Stack, unstack and reshape	64
2.5.3	Split, unsplit and unlist	66
2.5.4	Cut	66
2.5.5	Merge, union and intersect	68
2.5.6	<code>is.element()</code>	69
2.6	Manipulating strings	71
2.7	Assignments	72
<b>3</b>	<b>Presenting data</b>	<b>75</b>
3.1	Tables and the flavors of <code>apply()</code>	75
3.2	Bar plots	77
3.3	Histograms	81
3.4	Dot charts	85
3.5	Scatter plots	86
3.6	Lattice plots	88

3.7	Three-dimensional plots and contours	90
3.8	Assignments	90

## Part II Probability, densities and distributions

<b>4</b>	<b>Probability and random variables</b>	<b>97</b>
4.1	Set theory	98
4.1.1	Sets and algebra of sets	98
4.1.2	Set theory in $\mathbb{R}$	103
4.2	Trials, events and experiments	103
4.3	Definitions and properties of probability	108
4.3.1	Definitions of probability	108
4.3.2	Properties of probability	111
4.3.3	Equally likely events	112
4.3.4	Probability and set theory	112
4.4	Conditional probability and independence	113
4.4.1	Conditional probability	114
4.4.2	Independence	116
4.5	Algebra with probabilities	118
4.5.1	Sampling with and without replacement	118
4.5.2	Addition	119
4.5.3	Multiplication	120
4.5.4	Counting rules	120
4.6	Random variables	127
4.7	Assignments	128
<b>5</b>	<b>Discrete densities and distributions</b>	<b>137</b>
5.1	Densities	137
5.2	Distributions	141
5.3	Properties	143
5.3.1	Densities	144
5.3.2	Distributions	144
5.4	Expected values	144
5.5	Variance and standard deviation	146
5.6	The binomial	147
5.6.1	Expectation and variance	151
5.6.2	Decision making with the binomial	151
5.7	The Poisson	153
5.7.1	The Poisson approximation to the binomial	155
5.7.2	Expectation and variance	156
5.7.3	Variance of the Poisson density	157
5.8	Estimating parameters	161
5.9	Some useful discrete densities	163
5.9.1	Multinomial	163
5.9.2	Negative binomial	165
5.9.3	Hypergeometric	168
5.10	Assignments	171

<b>6</b>	<b>Continuous distributions and densities</b>	177
6.1	Distributions	177
6.2	Densities	180
6.3	Properties	181
6.3.1	Distributions	181
6.3.2	Densities	182
6.4	Expected values	183
6.5	Variance and standard deviation	184
6.6	Areas under density curves	185
6.7	Inverse distributions and simulations	187
6.8	Some useful continuous densities	189
6.8.1	Double exponential (Laplace)	189
6.8.2	Normal	191
6.8.3	$\chi^2$	193
6.8.4	Student- <i>t</i>	195
6.8.5	<i>F</i>	197
6.8.6	Lognormal	198
6.8.7	Gamma	199
6.8.8	Beta	201
6.9	Assignments	203
<b>7</b>	<b>The normal and sampling densities</b>	205
7.1	The normal density	205
7.1.1	The standard normal	207
7.1.2	Arbitrary normal	210
7.1.3	Expectation and variance of the normal	212
7.2	Applications of the normal	213
7.2.1	The normal approximation of discrete densities	214
7.2.2	Normal approximation to the binomial	215
7.2.3	The normal approximation to the Poisson	218
7.2.4	Testing for normality	220
7.3	Data transformations	225
7.4	Random samples and sampling densities	226
7.4.1	Random samples	227
7.4.2	Sampling densities	228
7.5	A detour: using R efficiently	230
7.5.1	Avoiding loops	230
7.5.2	Timing execution	230
7.6	The sampling density of the mean	232
7.6.1	The central limit theorem	232
7.6.2	The sampling density	232
7.6.3	Consequences of the central limit theorem	234
7.7	The sampling density of proportion	235
7.7.1	The sampling density	236
7.7.2	Consequence of the central limit theorem	238
7.8	The sampling density of intensity	239
7.8.1	The sampling density	239

7.8.2	Consequences of the central limit theorem	241
7.9	The sampling density of variance	241
7.10	Bootstrap: arbitrary parameters of arbitrary densities	242
7.11	Assignments	243

### Part III Statistics

<b>8</b>	<b>Exploratory data analysis</b>	251
8.1	Graphical methods	252
8.2	Numerical summaries	253
8.2.1	Measures of the center of the data	253
8.2.2	Measures of the spread of data	261
8.2.3	The Chebyshev and empirical rules	267
8.2.4	Measures of association between variables	269
8.3	Visual summaries	275
8.3.1	Box plots	275
8.3.2	Lag plots	276
8.4	Assignments	277
<b>9</b>	<b>Point and interval estimation</b>	283
9.1	Point estimation	284
9.1.1	Maximum likelihood estimators	284
9.1.2	Desired properties of point estimators	285
9.1.3	Point estimates for useful densities	288
9.1.4	Point estimate of population variance	292
9.1.5	Finding MLE numerically	293
9.2	Interval estimation	294
9.2.1	Large sample confidence intervals	295
9.2.2	Small sample confidence intervals	301
9.3	Point and interval estimation for arbitrary densities	304
9.4	Assignments	307
<b>10</b>	<b>Single sample hypotheses testing</b>	313
10.1	Null and alternative hypotheses	313
10.1.1	Formulating hypotheses	314
10.1.2	Types of errors in hypothesis testing	316
10.1.3	Choosing a significance level	317
10.2	Large sample hypothesis testing	318
10.2.1	Means	318
10.2.2	Proportions	323
10.2.3	Intensities	324
10.2.4	Common sense significance	325
10.3	Small sample hypotheses testing	326
10.3.1	Means	326
10.3.2	Proportions	327
10.3.3	Intensities	328

10.4	Arbitrary statistics of arbitrary densities	329
10.5	$p$ -values	330
10.6	Assignments	333
<b>11</b>	<b>Power and sample size for single samples</b>	<b>341</b>
11.1	Large sample	341
11.1.1	Means	342
11.1.2	Proportions	352
11.1.3	Intensities	356
11.2	Small samples	359
11.2.1	Means	359
11.2.2	Proportions	361
11.2.3	Intensities	363
11.3	Power and sample size for arbitrary densities	365
11.4	Assignments	365
<b>12</b>	<b>Two samples</b>	<b>369</b>
12.1	Large samples	370
12.1.1	Means	370
12.1.2	Proportions	375
12.1.3	Intensities	379
12.2	Small samples	380
12.2.1	Estimating variance and standard error	380
12.2.2	Hypothesis testing and confidence intervals for variance	382
12.2.3	Means	384
12.2.4	Proportions	386
12.2.5	Intensities	387
12.3	Unknown densities	388
12.3.1	Rank sum test	389
12.3.2	$t$ vs. rank sum	392
12.3.3	Signed rank test	392
12.3.4	Bootstrap	394
12.4	Assignments	396
<b>13</b>	<b>Power and sample size for two samples</b>	<b>401</b>
13.1	Two means from normal populations	401
13.1.1	Power	401
13.1.2	Sample size	404
13.2	Two proportions	406
13.2.1	Power	407
13.2.2	Sample size	409
13.3	Two rates	410
13.4	Assignments	415
<b>14</b>	<b>Simple linear regression</b>	<b>417</b>
14.1	Simple linear models	417
14.1.1	The regression line	418
14.1.2	Interpretation of simple linear models	419

14.2	Estimating regression coefficients	422
14.3	The model goodness of fit	428
14.3.1	The $F$ test	428
14.3.2	The correlation coefficient	433
14.3.3	The correlation coefficient vs. the slope	434
14.4	Hypothesis testing and confidence intervals	434
14.4.1	$t$ -test for model coefficients	435
14.4.2	Confidence intervals for model coefficients	435
14.4.3	Confidence intervals for model predictions	436
14.4.4	$t$ -test for the correlation coefficient	438
14.4.5	$z$ tests for the correlation coefficient	439
14.4.6	Confidence intervals for the correlation coefficient	441
14.5	Model assumptions	442
14.6	Model diagnostics	443
14.6.1	The hat matrix	445
14.6.2	Standardized residuals	447
14.6.3	Studentized residuals	448
14.6.4	The RSTUDENT residuals	449
14.6.5	The DFFITS residuals	453
14.6.6	The DFBETAS residuals	454
14.6.7	Cooke's distance	456
14.6.8	Conclusions	457
14.7	Power and sample size for the correlation coefficient	458
14.8	Assignments	459
<b>15</b>	<b>Analysis of variance</b>	<b>463</b>
15.1	One-way, fixed-effects ANOVA	463
15.1.1	The model and assumptions	464
15.1.2	The $F$ -test	469
15.1.3	Paired group comparisons	475
15.1.4	Comparing sets of groups	484
15.2	Non-parametric one-way ANOVA	488
15.2.1	The Kruskal-Wallis test	488
15.2.2	Multiple comparisons	491
15.3	One-way, random-effects ANOVA	492
15.4	Two-way ANOVA	495
15.4.1	Two-way, fixed-effects ANOVA	496
15.4.2	The model and assumptions	496
15.4.3	Hypothesis testing and the $F$ -test	500
15.5	Two-way linear mixed effects models	505
15.6	Assignments	509
<b>16</b>	<b>Simple logistic regression</b>	<b>511</b>
16.1	Simple binomial logistic regression	511
16.2	Fitting and selecting models	519
16.2.1	The log likelihood function	519
16.2.2	Standard errors of coefficients and predictions	521
16.2.3	Nested models	524

16.3	Assessing goodness of fit	525
16.3.1	The Pearson $\chi^2$ statistic	526
16.3.2	The deviance $\chi^2$ statistic	527
16.3.3	The group adjusted $\chi^2$ statistic	528
16.3.4	The ROC curve	529
16.4	Diagnostics	533
16.4.1	Analysis of residuals	533
16.4.2	Validation	536
16.4.3	Applications of simple logistic regression to $2 \times 2$ tables	536
16.5	Assignments	539
<b>17</b>	<b>Application: the shape of wars to come</b>	<b>541</b>
17.1	A statistical profile of the war in Iraq	541
17.1.1	Introduction	542
17.1.2	The data	542
17.1.3	Results	543
17.1.4	Conclusions	550
17.2	A statistical profile of the second Intifada	552
17.2.1	Introduction	552
17.2.2	The data	553
17.2.3	Results	553
17.2.4	Conclusions	561
	<b>References</b>	<b>563</b>
	<b>R Index</b>	<b>569</b>
	<b>General Index</b>	<b>583</b>