

# **Biostatistical Design and Analysis Using R**

## **A Practical Guide**

**Murray Logan**

 **WILEY-BLACKWELL**

A John Wiley & Sons, Inc., Publication

# Contents

<i>Preface</i>	xv
<i>R quick reference card</i>	xix
<i>General key to statistical methods</i>	xxvii
<b>1 Introduction to R</b>	<b>1</b>
1.1 Why R?	1
1.2 Installing R	2
1.2.1 Windows	2
1.2.2 Unix/Linux	2
1.2.3 MacOSX	3
1.3 The R environment	3
1.3.1 The console (command line)	4
1.4 Object names	4
1.5 Expressions, Assignment and Arithmetic	5
1.6 R Sessions and workspaces	6
1.6.1 Cleaning up	6
1.6.2 Workspaces	7
1.6.3 Current working directory	7
1.6.4 Quitting R	8
1.7 Getting help	8
1.8 Functions	9
1.9 Precedence	10
1.10 Vectors - variables	11
1.10.1 Regular or patterned sequences	12
1.10.2 Character vectors	13
1.10.3 Factors	15
1.11 Matrices, lists and data frames	16
1.11.1 Matrices	16
1.11.2 Lists	17
1.11.3 Data frames - data sets	18

1.12	Object information and conversion	18
1.12.1	Object information	18
1.12.2	Object conversion	20
1.13	Indexing vectors, matrices and lists	20
1.13.1	Vector indexing	21
1.13.2	Matrix indexing	22
1.13.3	List indexing	23
1.14	Pattern matching and replacement (character search and replace)	24
1.14.1	grep - pattern searching	24
1.14.2	regexpr - position and length of match	25
1.14.3	gsub - pattern replacement	26
1.15	Data manipulation	26
1.15.1	Sorting	26
1.15.2	Formatting data	27
1.16	Functions that perform other functions repeatedly	28
1.16.1	Along matrix margins	29
1.16.2	By factorial groups	30
1.16.3	By objects	30
1.17	Programming in R	30
1.17.1	Grouped expressions	31
1.17.2	Conditional execution – if and ifelse	31
1.17.3	Repeated execution – looping	32
1.17.4	Writing functions	34
1.18	An introduction to the R graphical environment	35
1.18.1	The plot() function	36
1.18.2	Graphical devices	39
1.18.3	Multiple graphics devices	40
1.19	Packages	42
1.19.1	Manual package management	42
1.19.2	Loading packages	45
1.20	Working with scripts	45
1.21	Citing R in publications	46
1.22	Further reading	47
<b>2</b>	<b>Data sets</b>	<b>48</b>
2.1	Constructing data frames	48
2.2	Reviewing a data frame - fix()	49
2.3	Importing (reading) data	50
2.3.1	Import from text file	50
2.3.2	Importing from the clipboard	51
2.3.3	Import from other software	51
2.4	Exporting (writing) data	52
2.5	Saving and loading of R objects	53
2.6	Data frame vectors	54
2.6.1	Factor levels	54

2.7	Manipulating data sets	56
2.7.1	Subsets of data frames – data frame indexing	56
2.7.2	The <code>%in%</code> matching operator	57
2.7.3	Pivot tables and aggregating datasets	58
2.7.4	Sorting datasets	58
2.7.5	Accessing and evaluating expressions within the context of a dataframe	59
2.7.6	Reshaping dataframes	59
2.8	Dummy data sets - generating random data	62
<b>3</b>	<b>Introductory statistical principles</b>	<b>65</b>
3.1	Distributions	66
3.1.1	The normal distribution	67
3.1.2	Log-normal distribution	68
3.2	Scale transformations	68
3.3	Measures of location	69
3.4	Measures of dispersion and variability	70
3.5	Measures of the precision of estimates - standard errors and confidence intervals	71
3.6	Degrees of freedom	73
3.7	Methods of estimation	73
3.7.1	Least squares (LS)	73
3.7.2	Maximum likelihood (ML)	74
3.8	Outliers	75
3.9	Further reading	75
<b>4</b>	<b>Sampling and experimental design with R</b>	<b>76</b>
4.1	Random sampling	76
4.2	Experimental design	83
4.2.1	Fully randomized treatment allocation	83
4.2.2	Randomized complete block treatment allocation	84
<b>5</b>	<b>Graphical data presentation</b>	<b>85</b>
5.1	The <code>plot()</code> function	86
5.1.1	The <code>type</code> parameter	86
5.1.2	The <code>xlim</code> and <code>ylim</code> parameters	87
5.1.3	The <code>xlab</code> and <code>ylab</code> parameters	88
5.1.4	The <code>axes</code> and <code>ann</code> parameters	88
5.1.5	The <code>log</code> parameter	88
5.2	Graphical Parameters	89
5.2.1	Plot dimensional and layout parameters	90
5.2.2	Axis characteristics	92
5.2.3	Character sizes	93
5.2.4	Line characteristics	93
5.2.5	Plotting character parameter - <code>pch</code>	93

5.2.6	Fonts	96
5.2.7	Text orientation and justification	98
5.2.8	Colors	98
5.3	Enhancing and customizing plots with low-level plotting functions	99
5.3.1	Adding points - <code>points()</code>	99
5.3.2	Adding text within a plot - <code>text()</code>	100
5.3.3	Adding text to plot margins - <code>mtext()</code>	101
5.3.4	Adding a legend - <code>legend()</code>	102
5.3.5	More advanced text formatting	104
5.3.6	Adding axes - <code>axis()</code>	107
5.3.7	Adding lines and shapes within a plot	108
5.4	Interactive graphics	113
5.4.1	Identifying points - <code>identify()</code>	113
5.4.2	Retrieving coordinates - <code>locator()</code>	114
5.5	Exporting graphics	114
5.5.1	Postscript - <code>postscript()</code> and <code>pdf()</code>	114
5.5.2	Bitmaps - <code>jpeg()</code> and <code>png()</code>	115
5.5.3	Copying devices - <code>dev.copy()</code>	115
5.6	Working with multiple graphical devices	115
5.7	High-level plotting functions for univariate (single variable) data	116
5.7.1	Histogram	116
5.7.2	Density functions	117
5.7.3	Q-Q plots	118
5.7.4	Boxplots	119
5.7.5	Rug charts	120
5.8	Presenting relationships	120
5.8.1	Scatterplots	120
5.9	Presenting grouped data	125
5.9.1	Boxplots	125
5.9.2	Boxplots for grouped means	125
5.9.3	Interaction plots - means plots	126
5.9.4	Bargraphs	127
5.9.5	Violin plots	128
5.10	Presenting categorical data	128
5.10.1	Mosaic plots	128
5.10.2	Association plots	129
5.11	Trellis graphics	129
5.11.1	<code>scales()</code> <i>parameters</i>	132
5.12	Further reading	133
<b>6</b>	<b>Simple hypothesis testing – one and two population tests</b>	<b>134</b>
6.1	Hypothesis testing	134
6.2	One- and two-tailed tests	136
6.3	<i>t</i> -tests	136

6.4	Assumptions	137
6.5	Statistical decision and power	137
6.6	Robust tests	139
6.7	Further reading	139
6.8	Key for simple hypothesis testing	140
6.9	Worked examples of real biological data sets	142
<b>7</b>	<b>Introduction to Linear models</b>	<b>151</b>
7.1	Linear models	152
7.2	Linear models in R	154
7.3	Estimating linear model parameters	156
7.3.1	Linear models with factorial variables	156
7.3.2	Linear model hypothesis testing	162
7.4	Comments about the importance of understanding the structure and parameterization of linear models	164
<b>8</b>	<b>Correlation and simple linear regression</b>	<b>167</b>
8.1	Correlation	168
8.1.1	Product moment correlation coefficient	169
8.1.2	Null hypothesis	169
8.1.3	Assumptions	169
8.1.4	Robust correlation	169
8.1.5	Confidence ellipses	170
8.2	Simple linear regression	170
8.2.1	Linear model	171
8.2.2	Null hypotheses	171
8.2.3	Assumptions	172
8.2.4	Multiple responses for each level of the predictor	173
8.2.5	Model I and II regression	173
8.2.6	Regression diagnostics	176
8.2.7	Robust regression	176
8.2.8	Power and sample size determination	177
8.3	Smoothers and local regression	178
8.4	Correlation and regression in R	178
8.5	Further reading	179
8.6	Key for correlation and regression	180
8.7	Worked examples of real biological data sets	184
<b>9</b>	<b>Multiple and curvilinear regression</b>	<b>208</b>
9.1	Multiple linear regression	208
9.2	Linear models	209
9.3	Null hypotheses	209
9.4	Assumptions	210
9.5	Curvilinear models	211
9.5.1	Polynomial regression	211

9.5.2 Nonlinear regression	214
9.5.3 Diagnostics	214
9.6 Robust regression	214
9.7 Model selection	214
9.7.1 Model averaging	215
9.7.2 Hierarchical partitioning	218
9.8 Regression trees	218
9.9 Further reading	219
9.10 Key and analysis sequence for multiple and complex regression	219
9.11 Worked examples of real biological data sets	224
<b>10 Single factor classification (ANOVA)</b>	<b>254</b>
10.0.1 Fixed versus random factors	254
10.1 Null hypotheses	255
10.2 Linear model	255
10.3 Analysis of variance	256
10.4 Assumptions	258
10.5 Robust classification (ANOVA)	259
10.6 Tests of trends and means comparisons	259
10.7 Power and sample size determination	261
10.8 ANOVA in R	261
10.9 Further reading	262
10.10 Key for single factor classification (ANOVA)	262
10.11 Worked examples of real biological data sets	265
<b>11 Nested ANOVA</b>	<b>283</b>
11.1 Linear models	284
11.2 Null hypotheses	285
11.2.1 <i>Factor A</i> - the main treatment effect	285
11.2.2 <i>Factor B</i> - the nested factor	285
11.3 Analysis of variance	286
11.4 Variance components	286
11.5 Assumptions	289
11.6 Pooling denominator terms	289
11.7 Unbalanced nested designs	290
11.8 Linear mixed effects models	290
11.9 Robust alternatives	292
11.10 Power and optimisation of resource allocation	292
11.11 Nested ANOVA in R	293
11.11.1 Error strata (aov)	293
11.11.2 Linear mixed effects models (lme and lmer)	294
11.12 Further reading	294
11.13 Key for nested ANOVA	294
11.14 Worked examples of real biological data sets	298

<b>12 Factorial ANOVA</b>	<b>313</b>
12.1 Linear models	314
12.2 Null hypotheses	314
12.2.1 Model 1 - fixed effects	315
12.2.2 Model 2 - random effects	316
12.2.3 Model 3 - mixed effects	317
12.3 Analysis of variance	317
12.3.1 Quasi <i>F</i> -ratios	320
12.3.2 Interactions and main effects tests	321
12.4 Assumptions	321
12.5 Planned and unplanned comparisons	321
12.6 Unbalanced designs	322
12.6.1 Missing observations	322
12.6.2 Missing combinations - missing cells	324
12.7 Robust factorial ANOVA	325
12.8 Power and sample sizes	327
12.9 Factorial ANOVA in R	327
12.10 Further reading	327
12.11 Key for factorial ANOVA	328
12.12 Worked examples of real biological data sets	334
<b>13 Unreplicated factorial designs – randomized block and simple repeated measures</b>	<b>360</b>
13.1 Linear models	363
13.2 Null hypotheses	363
13.2.1 <i>Factor A</i> - the main within block treatment effect	364
13.2.2 <i>Factor B</i> - the blocking factor	364
13.3 Analysis of variance	364
13.4 Assumptions	365
13.4.1 Sphericity	366
13.4.2 Block by treatment interactions	368
13.5 Specific comparisons	370
13.6 Unbalanced un-replicated factorial designs	370
13.7 Robust alternatives	371
13.8 Power and blocking efficiency	371
13.9 Unreplicated factorial ANOVA in R	371
13.10 Further reading	371
13.11 Key for randomized block and simple repeated measures ANOVA	372
13.12 Worked examples of real biological data sets	376
<b>14 Partly nested designs: split plot and complex repeated measures</b>	<b>399</b>
14.1 Null hypotheses	400
14.1.1 <i>Factor A</i> - the main between block treatment effect	400
14.1.2 <i>Factor B</i> - the blocking factor	401



14.1.3	<i>Factor C</i> - the main within-block treatment effect	401
14.1.4	<i>AC interaction</i> - the within block interaction effect	402
14.1.5	<i>BC interaction</i> - the within block interaction effect	402
14.2	Linear models	402
14.2.1	One between ( $\alpha$ ), one within ( $\gamma$ ) block effect	402
14.2.2	Two between ( $\alpha, \gamma$ ), one within ( $\delta$ ) block effect	402
14.2.3	One between ( $\alpha$ ), two within ( $\gamma, \delta$ ) block effects	403
14.3	Analysis of variance	403
14.4	Assumptions	403
14.5	Other issues	408
14.5.1	Robust alternatives	408
14.6	Further reading	408
14.7	Key for partly nested ANOVA	409
14.8	Worked examples of real biological data sets	413
<b>15</b>	<b>Analysis of covariance (ANCOVA)</b>	<b>448</b>
15.1	Null hypotheses	450
15.1.1	<i>Factor A</i> - the main treatment effect	450
15.1.2	<i>Factor B</i> - the covariate effect	450
15.2	Linear models	450
15.3	Analysis of variance	451
15.4	Assumptions	452
15.4.1	Homogeneity of slopes	453
15.4.2	Similar covariate ranges	454
15.5	Robust ANCOVA	455
15.6	Specific comparisons	455
15.7	Further reading	455
15.8	Key for ANCOVA	455
15.9	Worked examples of real biological data sets	457
<b>16</b>	<b>Simple Frequency Analysis</b>	<b>466</b>
16.1	The chi-square statistic	467
16.1.1	Assumptions	469
16.2	Goodness of fit tests	469
16.2.1	Homogeneous frequencies tests	469
16.2.2	Distributional conformity - Kolmogorov-Smirnov tests	469
16.3	Contingency tables	469
16.3.1	Odds ratios	470
16.3.2	Residuals	472
16.4	G-tests	472
16.5	Small sample sizes	473
16.6	Alternatives	474
16.7	Power analysis	474
16.8	Simple frequency analysis in R	475

16.9	Further reading	475
16.10	Key for Analysing frequencies	475
16.11	Worked examples of real biological data sets	477
<b>17</b>	<b>Generalized linear models (GLM)</b>	<b>483</b>
17.1	Dispersion (over or under)	485
17.2	Binary data - logistic (logit) regression	485
17.2.1	Logistic model	485
17.2.2	Null hypotheses	487
17.2.3	Analysis of deviance	488
17.2.4	Multiple logistic regression	488
17.3	Count data - Poisson generalized linear models	489
17.3.1	Poisson regression	489
17.3.2	Log-linear Modelling	489
17.4	Assumptions	492
17.5	Generalized additive models (GAM's) - non-parametric GLM	493
17.6	GLM and R	494
17.7	Further reading	495
17.8	Key for GLM	495
17.9	Worked examples of real biological data sets	498
	<i>Bibliography</i>	<b>531</b>
	<i>R index</i>	<b>535</b>
	<i>Statistics index</i>	<b>541</b>
Companion website for this book: <a href="http://wiley.com/go/logan/r">wiley.com/go/logan/r</a>		