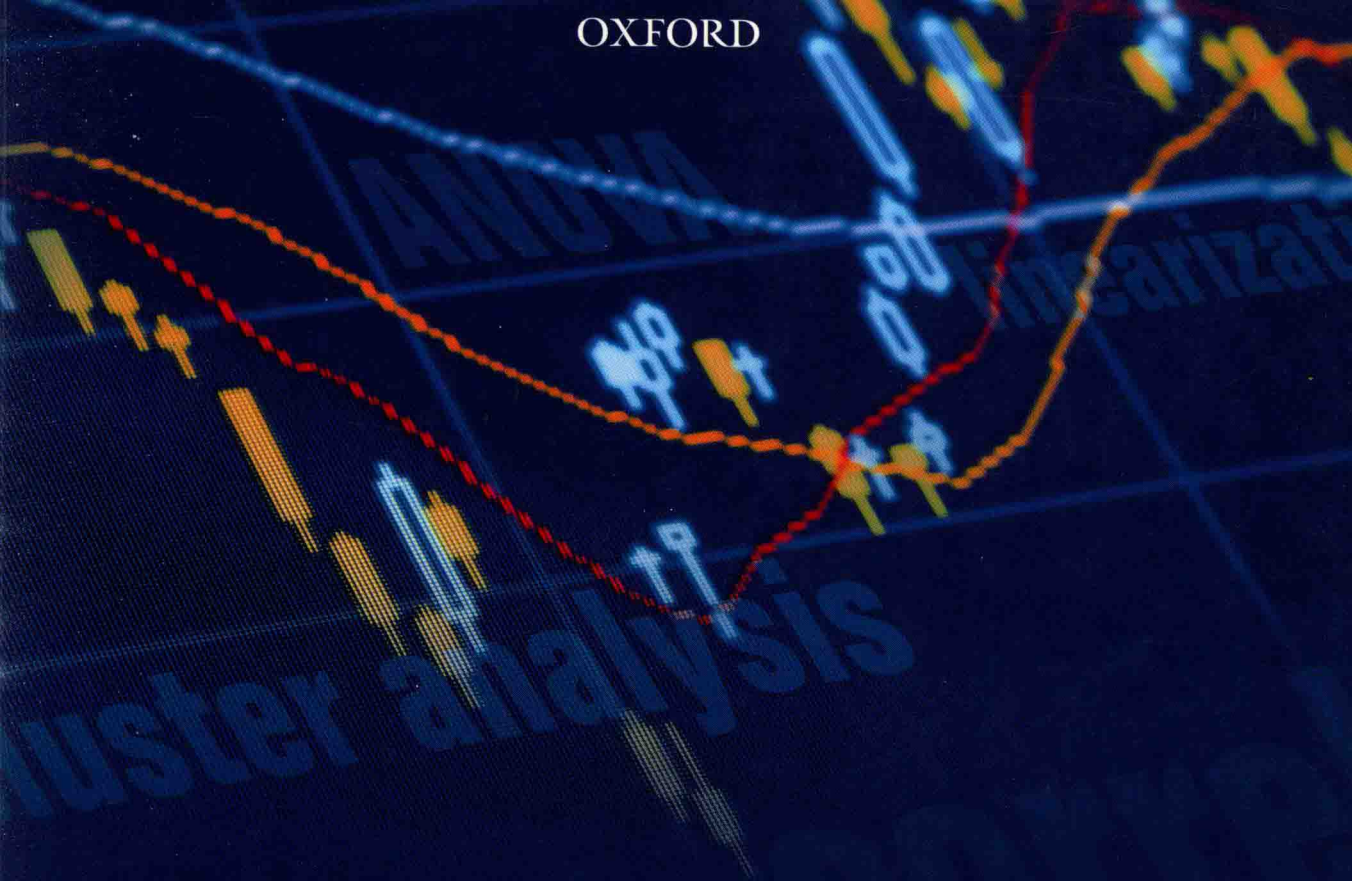
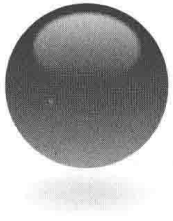


OXFORD



Graham Currell

scientific data analysis



Scientific Data Analysis

Graham Currell

Formerly University of the West of England, Bristol

OXFORD
UNIVERSITY PRESS

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Graham Currell 2015

The moral rights of the author have been asserted

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2014951237

ISBN 978-0-19-871254-1

Printed in Great Britain by Ashford Colour Press Ltd, Gosport, Hampshire

QR Code images are used throughout this book. QR Code is a registered trademark
of DENSO WAVE INCORPORATED. If your mobile device does not have a QR Code
reader try this website for advice www.mobile-barcodes.com/qr-code-software.

Excel is a registered trade mark of Microsoft Corporation. Microsoft product
screenshots reproduced with permission from Microsoft Corporation.

Portions of information contained in this publication/book are printed
with permission of Minitab Inc. All such material remains the exclusive
property and copyright of Minitab Inc. All rights reserved.

SPSS is a registered trade mark of IBM.

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.



Scientific Data Analysis

To my wife Jenny and son Felix for their continued support and encouragement.

About the author

Until his retirement in 2009, Graham Currell was a Principal Lecturer in physics at the University of the West of England, Bristol. During his early career, his particular interest was in the preparation of specialist training programmes to support staff in university science laboratories in Asia, the Middle East, Africa, and Central America but after 2000 he concentrated on the development of data analysis modules and self-study materials for science students, and from 2009 became a part-time research fellow, in which he further explored the development of online learning resources. *Scientific Data Analysis* builds on Graham's previous development of teaching materials for mathematics and statistics, including screen-capture videos in forensic, chemical, biological, and environmental science for the University of the West of England and the Royal Society of Chemistry. The approach reflects his extensive experience of providing tutorial and training support for students and staff carrying out research projects across both the physical and life sciences.

Welcome to *Scientific Data Analysis*

This book was written to satisfy the needs of three main target groups:

- Students following science degrees or masters courses who need support in understanding statistical data analysis, when encountered within taught courses and particularly when applied to their own developing experimental skills.
- University and college science staff wishing to reinforce or extend their own understanding of analytical techniques to help their supervision of student projects, particularly in topics on the border of their own specialisms.
- Teaching staff who are developing courses for science students and wish to structure the curriculum such that it prepares the students for handling the analysis of their own experimental project data.

Students typically learn the techniques of data analysis in two stages:

- Within taught modules during the first and/or second year of undergraduate studies.
- As part of a final year project, when faced with analysing their own experimental data.

This book is divided into two parts to reflect these two different approaches to learning. Part I, 'Understanding the statistics', develops the necessary statistical concepts with the *bottom-up* approach typically used in taught courses, and Part II, 'Analysing experimental data', starts from experimental data, and reviews, *top-down*, the possible techniques that could be used for their analysis.

The content and terminology used in Part I leads into the applications developed in Part II, and, in reverse, the techniques using Minitab and SPSS in Part II are supported through references to the basic statistical concepts in Part I.

For students using this book

If you, as a *first or second year* science student, are in the process of studying the general topic of data analysis, then Part I provides the core statistical concepts, which are developed in this book without complex mathematics but by using statistical models in Excel. The content reflects standard taught courses for science students, but also widens the range of techniques introduced in order to prepare you for the wide variety of different analytical problems encountered in final year projects.

If you, as a *final year* science student, are trying to find an analysis that is applicable to a set of your own experimental data, you should start in Part II by reading Chapter 5, from which you may move to one of Chapters 6 to 9, depending on the type of data involved. The 'Analytical options' feature in each section identifies possible analytical techniques that could be applied to the different types of data, and this approach helps you to select and use possible techniques relevant to your own particular data set. The content in Part II uses references to Part I to reinforce your understanding of the essential statistical concepts.

Video demonstrations

Short video clips are provided throughout the book to demonstrate the analyses using Minitab, SPSS and Excel. Together with the printed instructions, the videos will help you gain confidence in using the software and develop your experience for exploring menu options for other forms of analyses not directly covered in the book. The videos can be accessed directly by scanning the QR code images in the text with your smartphone. Alternatively you can view the videos from the links provided in the Online Resource Centre. Appendix I provides an index of the videos.

Case studies

There are a number of case studies throughout the book, most of which use related examples to develop a specific analytical theme. For example, it can be useful to see how the same or similar data can be analysed in different ways, and the case study links will take you to the relevant sections where the same case study continues. Each case study with more than one appearance starts with an overview which gives an outline of its theme and the locations in which each subsequent step can be found (not necessarily in the linear order of the book). Appendix II provides an index for the case studies.

Analytical software and data files

This book describes the use of Microsoft Excel 2013, Minitab 17, and IBM SPSS Statistics 22 for data analysis, giving the required keystrokes in the text and supported by videos accessible directly via the QR codes. The detailed steps described in Excel and SPSS are also the same as those used in the earlier versions of Excel 2010 and IBM SPSS Statistics 20 respectively. There are some differences between Minitab versions 16 and 17, many of which make little difference to the keystrokes required, except for the use of regression and the general linear models. Minitab 17 was introduced during the final stages of preparation of the book and the keystrokes and videos were updated to reflect the new software, but the legacy keystrokes and videos for Minitab 16 are still available in the Online Resource Centre. Some examples in the text were particularly relevant to Minitab 16 and these have been retained but clearly identified as such. Data files for the analyses in the book are also available for downloading from the Online Resource Centre.

Excel has widespread use for data handling in addition to its capabilities for statistical data analysis, and Minitab and SPSS have been chosen to demonstrate the 'next level' of analysis beyond Excel because of their easy-to-use, menu-driven operation. The approach developed using examples in Minitab and SPSS will support a greater understanding for solving problems if the student then moves on to using other packages, e.g. 'R', Prism.

Online Resource Centre

The Online Resource Centre can be found at <http://www.oxfordtextbooks.co.uk/orc/currell/>, and provides:

- Links to every video in the text, plus additional videos for Minitab 16.
- Download links for the modelling files developed in Excel, together with data files for Excel, Minitab, and SPSS.

In addition there is a 'new content' section which will be updated regularly to accommodate any new content or videos developed.

For the lecturer: about this book

This book evolved from the experience of working with very many students to support them analysing their own data within final year projects across a range of bioscience, forensic, environmental, and chemistry degree courses. It was instructive to discover the mismatch between the standard learning outcomes of first and second year 'statistics' modules and the practical problems faced by the students when they first encounter their own real world data analysis.

The ability to identify and implement the correct analytical technique requires both an *understanding* of the statistics involved together with the *experience* of a range of possible techniques. Unfortunately, for most science students, there is no simple linear path to achieving this combination, and they only develop a confident understanding of the statistics *after* having the experience of using it themselves to analyse their own real, and often somewhat scrappy, data. This book reflects this dichotomy, in that Part I provides a *bottom-up* review of the underlying statistics relevant to data analysis and Part II allows the student to address analytical problems, *top-down*, starting from the need to analyse typical science project data. A key aim of the book is to prepare the student for using his/her first 'solo' research project as an effective learning experience in data analysis, bringing together understanding of the statistics with its practical applications.

Part I of the book, 'Understanding the statistics', is targeted at the 'taught course' or 'fundamental concepts' phase of learning. The first two chapters develop the basics of experimental uncertainty and statistics within a strong scientific context. They also introduce terminology (e.g. sums of squares, confidence intervals, ANOVA tables) that links directly into the analytical techniques developed in Chapters 3 and 4. The aim of these chapters is then to expose the reader to a *wider range* of possible analytical approaches than is provided by most books concentrating on the standard *t*-test, ANOVA, and chi-squared analyses.

As examples, the first application of the *t*-statistic is not to develop the traditional *t*-test for mean values but to demonstrate its wider use by testing for a difference in the slopes of bacterial growth, and the 'repeated measures' ANOVA is not introduced in a questionnaire but used in a forensic test to differentiate between black inks.

Where it is important to understand the underlying statistics of the techniques, these are developed with a modelling approach using Excel supported by videos, which not only gives a more visual clarity to the analysis but also exposes the reader to wider possibilities in using Excel. With its student-centred approach, this book is an effective text/video resource that provides both content and context for learning the fundamental concepts of data analysis.

Part II, 'Analysing experimental data', is intended to be used mainly in a 'top down' approach to analysing experimental data, starting with Chapter 5 which introduces a phase of reflection to avoid rushing for the first analysis that will produce a (possibly irrelevant) result. The subsequent chapters and sections are then defined by the structure of the particular data set, allowing the student to investigate a wider set of analytical techniques than might have been considered initially.

The book prints keystroke instructions for SPSS and Minitab, together with a discussion of the resultant output, both of which are supported with step-by-step videos. Using this approach, the book provides self-study support for the individual reader, either student or lecturer, and would also be useful within the library of a statistics support centre for science students.

Contents

Part I **Understanding the statistics**

1	Statistical concepts	3
	Introduction	3
1.1	Data visualization	4
1.1.1	Graphical information	4
1.1.2	Boxplots	5
1.1.3	Raw data and calculated values	7
1.2	Scientific data	8
1.2.1	Experimental data	8
1.2.2	Data types	9
1.2.3	Type and value of data	10
1.3	Data distributions	10
1.3.1	Histogram	11
1.3.2	Distribution parameters	12
1.3.3	Standard distributions	14
1.4	Uncertainty and error	18
1.4.1	Error or uncertainty	18
1.4.2	True value	18
1.4.3	Experimental uncertainty	19
1.4.4	Combining uncertainties	20
1.4.5	Probability uncertainty	24
1.4.6	Identifying uncertainties	24
1.5	Sample data	27
1.5.1	Sample statistics	27
1.5.2	Confidence interval	29
1.5.3	Samples and populations	31
1.5.4	Known experimental uncertainty	35
1.5.5	Presenting results	36
1.6	Hypothesis tests	37
1.6.1	Test procedure	37
1.6.2	Hypothesis test and p -values	38
1.6.3	Errors in hypothesis tests	40
1.6.4	Bonferroni correction	41
2	Regression analysis	42
	Introduction	42
2.1	Regression statistics	43
2.1.1	Slope and intercept	43

2.1.2	ANOVA table	46
2.1.3	Correlation	48
2.1.4	Regression uncertainties	50
2.1.5	Quality of fit	51
2.2	Experimental uncertainties	53
2.2.1	Calibration uncertainty	53
2.2.2	Exact x/y intercepts	57
2.2.3	Known uncertainty	59
2.2.4	Weighting uncertainties	61
2.3	Linearization techniques	64
2.3.1	Change of variable	64
2.3.2	Using logarithms	66
2.3.3	Exponential relationships	67
2.3.4	Linearizing the exponential	68
2.3.5	Unknown power	71
2.3.6	Combined linearization	71
2.3.7	Error warning	72
2.4	Iteration using Solver	72
2.4.1	Operation of Solver	73
2.4.2	Maximum likelihood estimation	74
2.4.3	Nonlinear regression	75
3	Hypothesis testing	78
	Introduction	78
3.1	t -tests and z -tests	79
3.1.1	General principle of hypothesis testing	79
3.1.2	One sample t -test	81
3.1.3	Two sample t -test	83
3.1.4	Unequal variances	86
3.1.5	z -tests	86
3.2	Analysis of variance	87
3.2.1	F -test	87
3.2.2	Basic principle of ANOVA calculations	88
3.2.3	One-way ANOVA	89
3.2.4	Post hoc comparison tests	92
3.3	Multiple factors ANOVA	94
3.3.1	Two-way ANOVA	94
3.3.2	Interactions between the different factors	96
3.3.3	Analysis of covariance, ANCOVA	98
3.4	General linear model	101
3.4.1	General linear model	101
3.4.2	GLM, ANOVA, and the t -test	102
3.4.3	General regression	104
3.4.4	Fixed and random factor	106
3.4.5	Sequential and adjusted sums of squares	106
3.4.6	Lack of fit and error	109
3.4.7	Generalized linear model	109
3.5	Nonparametric analyses	111

3.5.1	Mann–Whitney example	111
3.5.2	Nonparametric and parametric test equivalents	113
3.6	Repeated measurements	114
3.6.1	Paired samples	115
3.6.2	Repeated measures	117
3.7	Chi-squared analyses	119
3.7.1	Tabulated data	120
3.7.2	One-way ‘goodness of fit’	120
3.7.3	Low value of chi-squared	123
3.7.4	Contingency table	123
3.7.5	Yates continuity correction	125
3.7.6	Likelihood ratio	126
3.7.7	Sample size limitations	126
3.8	Frequency and proportions	127
3.8.1	Probability distribution	127
3.8.2	One proportion test	127
3.8.3	Two proportions test	131
3.9	Resampling techniques	132
3.9.1	General approach to resampling	132
3.9.2	<i>t</i> -test and Mann–Whitney test	133
3.9.3	Chi-squared probabilities	136
4	Comparing data	140
	Introduction	140
4.1	Correlation	140
4.1.1	Linear correlation	140
4.1.2	Nonparametric correlation	143
4.1.3	Scientific context of correlation	146
4.1.4	Bivariate and partial correlation	146
4.2	Tests for association	148
4.2.1	Association and interaction	148
4.2.2	Tests for association	150
4.2.3	Fisher’s exact test	150
4.2.4	Linear by linear association	152
4.3	Strength of association	154
4.3.1	Association and agreement	154
4.3.2	Measures of association	155
4.3.3	Cramer’s <i>V</i> and <i>Phi</i>	156
4.3.4	Goodman and Kruskal’s <i>Lambda</i>	157
4.3.5	Concordance of data pairs	159
4.3.6	Nominal by interval association, <i>Eta</i>	160
4.4	Agreement between variables	162
4.4.1	R^2 goodness of fit	162
4.4.2	Agreement between two related variables	162
4.4.3	Agreement between several variables	166
4.4.4	Agreement within a contingency table	168
4.4.5	Binary agreement	171

Part II **Analysing experimental data**

5	Project data analysis	177
	Introduction	177
5.1	Preparing data for analysis	178
5.1.1	Case studies	178
5.1.2	Identifying the variables/factors	178
5.1.3	Understanding the uncertainty in the data	179
5.1.4	Scientific significance	180
5.1.5	Data entry into software	181
5.1.6	Reviewing data and objectives	183
5.2	Deriving test characteristics	185
5.2.1	Case studies	186
5.2.2	Beyond the exploratory phase	186
5.2.3	Selecting analyses	188
5.2.4	Combining data	189
5.2.5	Modelling response variables	190
5.3	Transforming and weighting data	193
5.3.1	Case studies	193
5.3.2	Software transformation	193
5.3.3	Common transformations	195
5.3.4	Weighting data	195
5.4	Normality and homoscedasticity	197
5.4.1	Case studies	197
5.4.2	Analytical approach	198
5.4.3	Anticipating normality	198
5.4.4	Differences in variance	199
5.4.5	Testing normality	199
5.4.6	Using residuals	202
5.4.7	Data transformations	205
6	Single response variable	209
	Introduction	209
6.1	One sample	209
6.1.1	Example data	209
6.1.2	Analytical options	210
6.1.3	Describing the data	211
6.1.4	One sample <i>t</i> -test	214
6.1.5	Wilcoxon test	215
6.1.6	SPSS nonparametric tests	216
6.1.7	Proportions	217
6.2	Two samples	218
6.2.1	Example data	218
6.2.2	Analytical options	220
6.2.3	Describing the data	221
6.2.4	Comparing variances	222
6.2.5	Two sample <i>t</i> -test	222
6.2.6	Nonparametric tests	223

6.2.7 Paired t-test	225
6.2.8 Paired Wilcoxon test	225
6.2.9 Unrelated binary data	226
6.3 One factor	227
6.3.1 Example data	227
6.3.2 Analytical options	229
6.3.3 Describing the data	229
6.3.4 Normality and equality of variance (homoscedasticity)	230
6.3.5 GLM/ANOVA	231
6.3.6 Post hoc comparison tests	233
6.3.7 Kruskal–Wallis test	234
6.3.8 Repeated measures	235
6.4 Multiple factors and interactions	236
6.4.1 Example data	236
6.4.2 Analytical options	239
6.4.3 Describing the data	239
6.4.4 GLM/ANOVA	241
6.4.5 Checking for normality and homoscedasticity	243
6.4.6 Nonparametric ANOVAs	245
6.4.7 Generalized linear model	246
6.4.8 Analysis of covariance, ANCOVA	247
7 Related variables	249
Introduction	249
7.1 Regression, correlation, and agreement	249
7.1.1 Example data	250
7.1.2 Analytical options	251
7.1.3 Describing the data	252
7.1.4 Correlation	253
7.1.5 Linear regression and calibration	254
7.1.6 Agreement between results	256
7.2 Nonlinear relationships	257
7.2.1 Example data	257
7.2.2 Analytical options	258
7.2.3 Iterative nonlinear regression	258
7.2.4 Deriving the mathematical model	261
7.2.5 General regression	262
7.3 General x–y data	264
7.3.1 Example data	264
7.3.2 Analytical options	265
7.3.3 Identifying relevant analytical characteristics	265
7.3.4 Describing the data	266
7.3.5 Smoothing convolutes	267
7.3.6 Differentiating convolutes	270
7.3.7 Spectral analysis	270
8 Frequency data	274
Introduction	274
8.1 Single variable	274

8.1.1	Example data	274
8.1.2	Analytical options	276
8.1.3	Describing categorical data	276
8.1.4	Editing histograms	278
8.1.5	Chi-squared 'goodness of fit' test	279
8.1.6	Testing distributions	281
8.1.7	Tabulation of data	282
8.1.8	Binning	283
8.2	Contingency tables	283
8.2.1	Example data	284
8.2.2	Analytical options	285
8.2.3	Describing the data	286
8.2.4	Contingency tables and cross-tabulation	287
8.2.5	Progression within the table	290
8.2.6	Data consolidation	291
8.2.7	Low expected frequencies	292
8.2.8	Layered contingency tables	293
8.3	Binary output data	294
8.3.1	Example data	294
8.3.2	Analytical options	295
8.3.3	Logit and probit linearization	295
8.3.4	Binary regression	298
8.3.5	Binary probabilities and ROC plots	300
9	Multiple variables	304
	Introduction	304
9.1	Modelling multiple variables	304
9.1.1	Example data	304
9.1.2	Analytical options	305
9.1.3	Cluster analysis	306
9.1.4	Principal component analysis	308
9.1.5	Factor analysis	311
9.1.6	Multiple regression	312
9.2	Multiple questions	315
9.2.1	Example data	315
9.2.2	Describing the data	317
9.2.3	Testing for normality and homoscedasticity	318
9.2.4	Analysing an individual variable	319
9.2.5	Dependence of specific factors	319
9.2.6	Comparing variables as unrelated data	320
9.2.7	Modelling interrelated variables	320
9.2.8	Comparing related variables	321
9.2.9	Ordinal responses	322
9.2.10	Multiple variables	322
	Appendix I Videos available in the Online Resource Centre	325
	Appendix II Case studies used throughout this book	328
	Index	331

Part I

Understanding the statistics

Part I approaches an understanding of analytical techniques in science from a *statistical* perspective. It develops an understanding of how key analytical techniques work and the scientific interpretation of their results. With this approach, it supports first and second year modules in statistical data analysis, but also acts as a reference resource for students subsequently meeting a technique for the first time. The implementation of many of the analytical techniques, developed in Part I, is then described in Part II from a *scientific* perspective using SPSS and Minitab.

Chapter 1. Statistical concepts provides the key topics and statistics that underpin the analytical techniques developed in subsequent chapters. The content reflects the standard elements of an introductory course in statistics, but the approach and terminology is designed to link into later applications.

Chapter 2. Regression analysis builds on the familiar 'best-fit straight line' analysis as an introduction to important analytical techniques. It provides an understanding of its practical implementation in experimental science using Excel, and develops the approach and terminology used in Minitab and SPSS, leading to the introduction of general linear models of analysis.

Chapter 3. Hypothesis testing provides an understanding of the process and significance of hypothesis testing in science, covering a wide range of underlying parametric and nonparametric techniques, from t-tests to Monte Carlo re-sampling. The specific concepts are developed, not through extensive statistical theory, but through the use of modelling in Excel, which provides a more relevant perspective for most science students, coupled with (possibly) new skills in Excel.

Chapter 4. Comparing data considers a range of analytical techniques that are often neglected in teaching but do address important questions in science. These relate to the strengths of agreement, association and interaction between the factors and variables in a scientific system.

