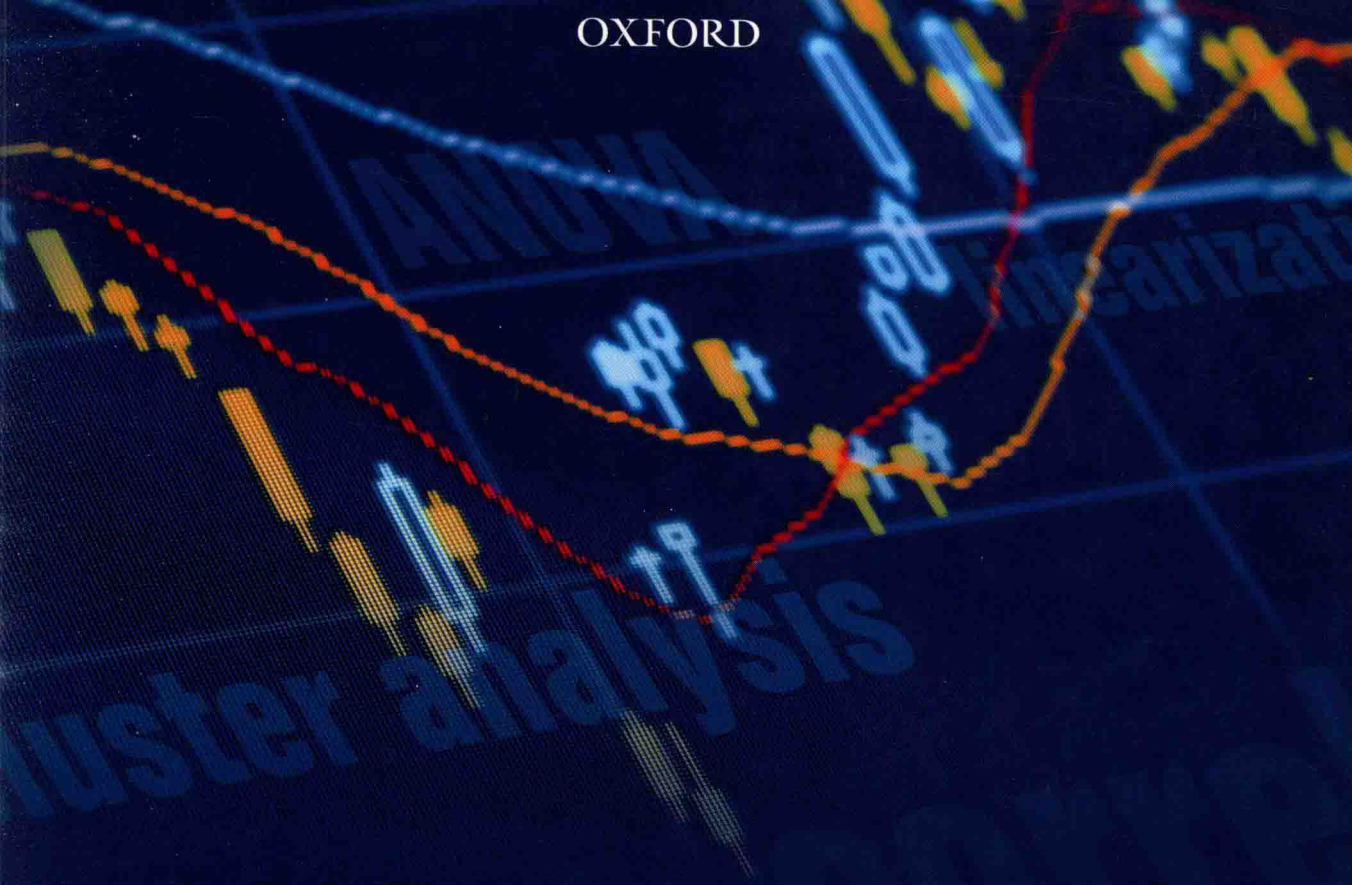
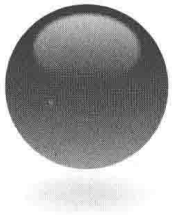
OXFORD



Graham Currell

scientific data analysis

online resource centre



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

Impression: 1

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

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 II **Analysing experimental data**

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.