SCIENTIFIC INFERENCE

Providing the knowledge and practical experience to begin analysing scientific data, this book is ideal for physical sciences students wishing to improve their data handling skills.

The book focuses on explaining and developing the practice and understanding of basic statistical analysis, concentrating on a few core ideas, such as the visual display of information, modelling using the likelihood function, and simulating random data.

Key concepts are developed through a combination of graphical explanations, worked examples, example computer code and case studies using real data. Students will develop an understanding of the ideas behind statistical methods and gain experience in applying them in practice. Further resources are available at www.cambridge.org/9781107607590, including data files for the case studies so students can practice analysing data, and exercises to test students' understanding.

SIMON VAUGHAN is a Reader in the Department of Physics and Astronomy, University of Leicester, where he has developed and runs a highly regarded course for final year physics students on the subject of statistics and data analysis.

SCIENTIFIC INFERENCE

Learning from data

SIMON VAUGHAN
University of Leicester

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107024823

First published 2013

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Vaughan, Simon, 1976– author.
Scientific inference : learning from data / Simon Vaughan.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-02482-3 (hardback) – ISBN 978-1-107-60759-0 (paperback)
1. Mathematical statistics – Textbooks. I. Title.
QA276.V34 2013
519.5 – dc23 2013021427

ISBN 978-1-107-02482-3 Hardback
ISBN 978-1-107-60759-0 Paperback

For my family

Contents

For the student

Science is not about certainty, it is about dealing rigorously with uncertainty. The tools for this are statistical. Statistics and data analysis are therefore an essential part of the scientific method and modern scientific practice, yet most students of physical science get little explicit training in statistical practice beyond basic error handling. The aim of this book is to provide the student with both the knowledge and the practical experience to begin analysing new scientific data, to allow progress to more advanced methods and to gain a more statistically literate approach to interpreting the constant flow of data provided by modern life.

More specifically, if you work through the book you should be able to accomplish the following.

- Explain aspects of the scientific method, types of logical reasoning and data analysis, and be able to critically analyse statistical and scientific arguments.
- Calculate and interpret common quantitative and graphical statistical summaries.
- Use and interpret the results of common statistical tests for difference and association, and straight line fitting.
- Use the calculus of probability to manipulate basic probability functions.
- Apply and interpret model fitting, using e.g. least squares, maximum likelihood.
- Evaluate and interpret confidence intervals and significance tests.

Students have asked me whether this is a book about statistics or data analysis or statistical computing. My answer is that they are so closely connected it is difficult to untangle them, and so this book covers areas of all three.

The skills and arguments discussed in the book are highly transferable: statistical presentations of data are used throughout science, business, medicine, politics and the news media. An awareness of the basic methods involved will better enable you to use and critically analyse such presentations – this is sometimes called *statistical literacy*.

In order to understand the book, you need to be familiar with the mathematical methods usually taught in the first year of a physics, engineering or chemistry degree (differential and integral calculus, basic matrix algebra), but this book is designed so that the probability and statistics content is entirely self-contained.

For the instructor

This book was written because I could not find a suitable textbook to use as the basis of an undergraduate course on scientific inference, statistics and data analysis. Although there are good books on different aspects of introductory statistics, those intended for physicists seem to target a post-graduate audience and cover either too much material or too much detail for an undergraduate-level first course. By contrast, the 'Intro to stats' books aimed at a broader audience (e.g. biologists, social scientists, medics) tend to cover topics that are not so directly applicable for physical scientists. And the books aimed at mathematics students are usually written in a style that is inaccessible to most physics students, or in a recipe-book style (aimed at science students) that provides ready-made solutions to common problems but develops little understanding along the way.

This book is different. It focuses on explaining and developing the *practice* and *understanding* of basic statistical analysis, concentrating on a few core ideas that underpin statistical and data analysis, such as the visual display of information, modelling using the likelihood function, and simulating random data. Key concepts are developed using several approaches: verbal exposition in the main text, graphical explanations, case studies drawn from some of history's great physics experiments, and example computer code to perform the necessary calculations.[1] The result is that, after following all these approaches, the student should both understand the ideas behind statistical methods and have experience in applying them in practice.

The book is intended for use as a textbook for an introductory course on data analysis and statistics (with a bias towards students in physics) or as self-study companion for professionals and graduate students. The book assumes familiarity with calculus and linear algebra, but no previous exposure to probability or statistics

[1] These are based on R, a freely available software package for data analysis and statistics and used in many statistics textbooks.

is assumed. It is suitable for a wide range of undergraduate and postgraduate science students.

The book has been designed with several special features to improve its value and effectiveness with students:

- several complete data analysis case studies using real data from some of history's great experiments
- 'example boxes' – approximately 20 boxes throughout the text that give specific, worked examples for concepts as they are discussed
- 'computer practice boxes' – approximately 90 boxes throughout the text that give working R code to perform the calculations discussed in the text or produce the plots shown
- graphical explanations of important concepts
- appendices that provide technical details supplementary to the main text
- a well-populated glossary of terms and list of notational conventions.

The emphasis on a few core ideas and their practical applications means that some subjects usually covered in introductory statistics texts are given little or no treatment here. Rigorous mathematical proofs are not covered – the interested reader can easily consult any good reference work on probability theory or mathematical statistics to check these. In addition, we do not cover some topics of 'classical' statistics that are dealt with in other introductory works. These topics include

- more advanced distribution functions (beta, gamma, multinomial, . . .)
- ANOVA and the generalised linear model
- characteristic functions and the theory of moments
- decision and information theories
- non-parametric tests
- experimental design
- time series analysis
- multivariate analysis (principal components, clustering, . . .)
- survival analysis
- spatial data analysis.

Upon completion of this book the student should be in a much better position to understand any of these topics from any number of more advanced or comprehensive texts.

Perhaps the 'elephant in the room' question is: what about Bayesian methods? Unfortunately, owing to practical limitations there was not room to include full chapters developing Bayesian methods. I hope I have designed the book in such a way that it is not wholly frequentist or Bayesian. The emphasis on model fitting

using the likelihood function (Chapter 6) could be seen as the first step towards a Bayesian analysis (i.e. implicitly using flat priors and working towards the posterior mode). Fortunately, there are many good books on Bayesian data analysis that can then be used to develop Bayesian ideas explicitly. I would recommend Gelman *et al*. (2003) generally and Sivia and Skilling (2006) or Gregory (2005) for physicists in particular. Albert (2007) also gives a nice 'learn as you compute' introduction to Bayesian methods using R.