STATISTICAL DATA ANALYSIS FOR THE PHYSICAL SCIENCES

Data analysis lies at the heart of every experimental science. Providing a modern introduction to statistics, this book is ideal for undergraduates in physics. It introduces the necessary tools required to analyse data from experiments across a range of areas, making it a valuable resource for students.

In addition to covering the basic topics, the book also takes in advanced and modern subjects, such as neural networks, decision trees, fitting techniques and issues concerning limit or interval setting. Worked examples and case studies illustrate the techniques presented, and end-of-chapter exercises help test the reader's understanding of the material.

ADRIAN BEVAN is a Reader in Particle Physics in the School of Physics and Astronomy, Queen Mary, University of London. He is an expert in quark flavour physics and has been analysing experimental data for over 15 years.

STATISTICAL DATA ANALYSIS FOR THE PHYSICAL SCIENCES

ADRIAN BEVAN
Queen Mary, University of London

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9781107670341

First published 2013

Printed and bound in the United Kingdom by the MPG Books Group

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

ISBN 978-1-107-03001-5 Hardback
ISBN 978-1-107-67034-1 Paperback

Contents

Preface

The foundations of science are built upon centuries of careful observation. These constitute measurements that are interpreted in terms of hypotheses, models, and ultimately well-tested theories that may stand the test of time for only a few years or for centuries. In order to understand what a single measurement means we need to appreciate a diverse range of statistical methods. Without such an appreciation it would be impossible for scientific method to turn observations of nature into theories that describe the behaviour of the Universe from sub-atomic to cosmic scales. In other words science would be impracticable without statistical data analysis. The data analysis principles underpinning scientific method pervade our everyday lives, from the use of statistics we are subjected to through advertising to the smooth operation of SPAM filters that we take for granted as we read our e-mail. These methods also impact upon the wider economy, as some areas of the financial industry use data mining and other statistical techniques to predict trading performance or to perform risk analysis for insurance purposes.

This book evolved from a one-semester advanced undergraduate course on statistical data analysis for physics students at Queen Mary, University of London with the aim of covering the rudimentary techniques required for many disciplines, as well as some of the more advanced topics that can be employed when dealing with limited data samples. This has been written by a physicist with a non-specialist audience in mind. This is not a statistics book for statisticians, and references have been provided for the interested reader to refer to for more rigorous treatment of the techniques discussed here. As a result this book provides an up-to-date introduction to a wide range of methods and concepts that are needed in order to analyse data. Thus this book is a mixture of a traditional text book approach and a teach by example approach. By providing these opposing viewpoints it is hoped that the reader will find the material more accessible. Throughout the book, a number of case studies are presented with possible solutions discussed in detail. The purpose of these sections is to consolidate the more abstract notions discussed in the book and apply them to an example. In some instances the case study may appear somewhat abstract and specific to scientific research; however, where possible more widely applicable problems have been included. At the end of each chapter there is a summary of the main issues raised, followed by a number of example questions to help the reader practise and gain a deeper understanding of the material included. Solutions to questions are presented at the end of the book.

The Introduction motivates the importance of studying statistical methods when analysing data by looking at three common problems encountered early within the life of a physicist: measuring g, testing Ohm's law and studying the law of radioactive decay. Following this motivational introduction the book is divided into two parts: (i) the foundations of statistical data analysis from set notation through to confidence intervals, and (ii) discussion of more advanced topics in the form of optimisation, fitting, and data mining. The material in the first part of the book is ordered logically so that successive sections build on material discussed in the earlier ones, while the second part of the book contains stand alone chapters that depend on concepts developed in the first part. These later chapters can be read in any order.

The first part of this book starts with an introduction to sets and Venn diagrams that provide some of the language that we use to discuss data. Having developed this language, the concept of probability is formally introduced in Chapter 3. Readers who are familiar with these concepts already may wish to skip over the first two chapters and proceed straight to the discussion in Chapter 4 on how to visualise and quantify data. Distributions of data are often described by simple functions that are used to represent the probability of observing data with a certain value. A number of useful distributions are described in Chapter 5, and Appendix B builds on this topic by discussing a number of additional functions that may be of use. Measurements are based on the determination of some central value of an observable quantity, with an uncertainty or error on that observable. Issues surrounding uncertainties and errors are introduced in Chapter 6, and this topic is further developed in Chapter 7. Chapter 8 discusses hypothesis testing and brings together many of the earlier concepts in the book.

The second part of the book presents more advanced topics. Chapter 9 discusses fitting data given some assumed model using χ^2 and likelihood methods. This relies heavily on concepts developed in Chapters 5 and 6, and Appendix B. Chapter 10 discusses data mining, or how to efficiently separate two classes of data, for example signal from background using numerical methods. The methods discussed include the use of 'cut-based' selection, the Bayesian classifier, Fisher's linear discriminant, artificial neural networks, and decision trees.

To avoid interrupting the flow of the text, a number of detailed appendices have been prepared. The most important of these appendices is a collection of probability tables, which is conveniently located at the end of the book in order to provide a quick reference to the reader. There is also a glossary of terms intended to help the reader when referring back to the book some time after an initial reading. Appendices listing a number of commonly used probability density functions, and elementary numerical integration techniques have also been provided. While these are not strictly required in order to understand the concepts introduced in the book, they have been included in order to make this a more complete resource for readers who wish to study this topic beyond an undergraduate course.

There are a number of technical terms introduced throughout this book. When a new term is introduced, that term is highlighted in ***bold-italic text*** to help the reader refer back to this description at a later time.

I would like to thank colleagues who have provided me with feedback on the draft of this book, and in particular Peter Crew.