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
# Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td></td>
<td>ix</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Measuring $g$, the coefficient of acceleration due to gravity</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Verification of Ohm’s law</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Measuring the half-life of an isotope</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>Summary</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Sets</td>
<td>12</td>
</tr>
<tr>
<td>2.1</td>
<td>Relationships between sets</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Summary</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>Probability</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Elementary rules</td>
<td>21</td>
</tr>
<tr>
<td>3.2</td>
<td>Bayesian probability</td>
<td>21</td>
</tr>
<tr>
<td>3.3</td>
<td>Classic approach</td>
<td>24</td>
</tr>
<tr>
<td>3.4</td>
<td>Frequentist probability</td>
<td>25</td>
</tr>
<tr>
<td>3.5</td>
<td>Probability density functions</td>
<td>26</td>
</tr>
<tr>
<td>3.6</td>
<td>Likelihood</td>
<td>27</td>
</tr>
<tr>
<td>3.7</td>
<td>Case studies</td>
<td>27</td>
</tr>
<tr>
<td>3.8</td>
<td>Summary</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>Visualising and quantifying the properties of data</td>
<td>35</td>
</tr>
<tr>
<td>4.1</td>
<td>Visual representation of data</td>
<td>35</td>
</tr>
<tr>
<td>4.2</td>
<td>Mode, median, mean</td>
<td>37</td>
</tr>
<tr>
<td>4.3</td>
<td>Quantifying the spread of data</td>
<td>39</td>
</tr>
<tr>
<td>4.4</td>
<td>Presenting a measurement</td>
<td>41</td>
</tr>
<tr>
<td>4.5</td>
<td>Skew</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>v</strong></td>
</tr>
</tbody>
</table>
## Contents

4.6 Measurements of more than one observable 44  
4.7 Case study 52  
4.8 Summary 53  
Exercises 53  

5 Useful distributions 56  
5.1 Expectation values of probability density functions 57  
5.2 Binomial distribution 57  
5.3 Poisson distribution 62  
5.4 Gaussian distribution 65  
5.5 $\chi^2$ distribution 67  
5.6 Computational issues 68  
5.7 Summary 70  
Exercises 70  

6 Uncertainty and errors 72  
6.1 The nature of errors 72  
6.2 Combination of errors 75  
6.3 Binomial error 79  
6.4 Averaging results 81  
6.5 Systematic errors and systematic bias 82  
6.6 Blind analysis technique 84  
6.7 Case studies 85  
6.8 Summary 90  
Exercises 91  

7 Confidence intervals 93  
7.1 Two-sided intervals 93  
7.2 Upper and lower limit calculations 94  
7.3 Limits for a Gaussian distribution 96  
7.4 Limits for a Poisson distribution 98  
7.5 Limits for a binomial distribution 100  
7.6 Unified approach to analysis of small signals 101  
7.7 Monte Carlo method 105  
7.8 Case studies 106  
7.9 Summary 111  
Exercises 112  

8 Hypothesis testing 114  
8.1 Formulating a hypothesis 114  
8.2 Testing if the hypothesis agrees with data 115  
8.3 Testing if the hypothesis disagrees with data 117
Contents

8.4 Hypothesis comparison 117
8.5 Testing the compatibility of results 119
8.6 Establishing evidence for, or observing a new effect 120
8.7 Case studies 124
8.8 Summary 125
Exercises 126

9 Fitting 128
9.1 Optimisation 128
9.2 The least squares or $\chi^2$ fit 131
9.3 Linear least-squares fit 134
9.4 Maximum-likelihood fit 136
9.5 Combination of results 140
9.6 Template fitting 142
9.7 Case studies 142
9.8 Summary 150
Exercises 151

10 Multivariate analysis 153
10.1 Cutting on variables 154
10.2 Bayesian classifier 157
10.3 Fisher discriminant 158
10.4 Artificial neural networks 162
10.5 Decision trees 169
10.6 Choosing an MVA technique 171
10.7 Case studies 174
10.8 Summary 177
Exercises 178

Appendix A Glossary 181
Appendix B Probability density functions 186
Appendix C Numerical integration methods 198
Appendix D Solutions 201
Appendix E Reference tables 207

References 216
Index 218
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
Preface

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 $\chi^2$ 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 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.