Preface

Science is fundamentally about learning from data, and doing so in the presence of uncertainty. Uncertainty arises inevitably and avoidably in many guises. It comes from noise in our measurements: we cannot measure exactly. It comes from sampling effects: we cannot measure everything. It comes from complexity: data may be numerous, high dimensional, and correlated, making it difficult to see structures.

This book is an introduction to statistical methods for analysing data. It presents the major concepts of probability and statistics as well as the computational tools we need to extract meaning from data in the presence of uncertainty.

Just as science is about learning from data, so learning from data is nearly synonymous with data modelling. This is because once we have a set of data, we normally want to identify its underlying structure, something we invariably represent with a model. Fitting and comparing models is therefore one of the cornerstones of statistical data analysis. This process of obtaining meaning from data and reasoning is what is meant by *inference*.

Alas, statistics is all too often taught as a set of seemingly unconnected, ad hoc recipes. Having identified what appears to be a relevant statistical test from a menu – according to the properties of your data and your assumptions – you then apply a procedure that delivers some numerical measure of significance. This kind of approach does little to promote your confidence in the result, and it will leave you lost when your assumptions aren't on the menu. My goal in this book is to show that the process of analysing and interpreting data can be done within a simple probabilistic framework. Probability is central to interpreting data because it quantifies both the uncertainty in the data and our confidence in models derived from them. I will show that there are basic principles for tackling problems that are built on a solid probabilistic foundation. Armed with this know-how you will be able to apply these principles and methods to a wide range of data analysis problems beyond the scope of this book.

This book is aimed primarily at undergraduate and graduate science students. Knowledge of calculus is assumed, but no specific experience with probability or statistics is required. My emphasis is on the concepts and on illustrating them with examples, using both analytical and numerical methods. This is not a maths book, and while I hope I am reasonably rigorous, I omit some more complex derivations and formal proofs. I also hope this book will be useful for more experienced practitioners – in particular those with limited or no exposure to Bayesian methods – by providing an overview of the main concepts and techniques.

Many real-world problems in data analysis must be solved with a computer. One of my goals is to show how theoretical ideas can be converted into practical solutions. To support this I include code in the R language (`https://www.r-project.org`). R is easy to learn and use, has a convenient environment for plotting, and is widely used within and outside academia. I encourage you to use and modify the code provided, for which some basic knowledge of R will be necessary. This is not a guide to doing statistics in R, however, and you can follow everything in the book without actually using the programs.

Before delving in, let me give you a roadmap. The first chapter covers the basic ideas of probability and distributions. Chapter 2 looks at the concepts of estimation, measurement, and uncertainty. Together, these two chapters provide essential background material for what follows. Chapter 3 is an introduction to probabilistic inference, for which I use simple problems to show how the measured data and prior information can be, and must be, combined. This chapter introduces many of the concepts in inference that are expanded upon later in the book. Chapter 4 looks at least squares and maximum likelihood in the context of curve fitting. This is a straightfoward problem, but illustrates many of the concepts needed to tackle more complex problems. Chapters 5 and 6 look at parameter estimation for one and two parameter problems respectively, by calculating the posterior probability density over the parameters. Some problems have analytic solutions based on so-called conjugate priors or the use of standard integrals. In these chapters I will also discuss the issue of assigning priors and how to summarize distributions. In chapter 7 I describe a method for approximating the posterior, and I will introduce the method of density estimation, which we use to describe distributions we have sampled from. Whereas low dimensional posteriors can often be found by brute force evaluation on a grid, higher dimensional problems are better solved by sampling from the distribution. Chapter 8 introduces Monte Carlo methods for doing sampling; we look in particular at the Metropolis method. This can also be used for integration, so here I also summarize the uses of integration in inference. Chapter 9 applies the Metropolis method to find the posterior probability density function for higher dimensional problems, again taking curve fitting problems as the examples. Here I will further investigate what to do (remaining in a probabilistic framework) when we have outliers or errors on both axes. In chapter 10 I look at frequentist hypothesis testing and discuss its uses and weaknesses. This will be contrasted in chapter 11 with the Bayesian approach to model comparison using the marginal likelihood, where I will also mention some alternative metrics of model suitability. In the final chapter I look at some of the techniques available for tackling more complicated problems. These include cross-validation, regularization, basis functions, and bootstrapping.

Readers who are familiar with basic statistical concepts and who only want an introduction to probabilistic (Bayesian) methods can read chapters 3, 5, 6, 9, and 11, as well as chapter 8 if you are not familiar with Monte Carlo methods.

Most of the notation I use should be clear from the context, but here is a quick summary of my main conventions. Where relevant to define a base I use $\log$ for base 10 and $\ln$ for the natural logarithm. (Annoyingly in R, these are `log10` and `log` respectively.) Lower case letters can indicate scalars or vectors. Where necessary to make a distinction, as in matrix algebra, I use bold face for vectors. Matrices are written with upper case letters. An $N \times J$ matrix has N rows and J columns. A vector with J elements is by default a column vector, i.e. $J \times 1$. The transpose is indicated with $^{\mathsf{T}}$ as in X^{T}. Curly brackets are used to indicate a sample, so $\{x\}$ is a set of values of x. $P(x)$ usually indicates a probability density function (PDF) over x. In the few places where we deal with discrete distributions, $P(x)$ will refer to an actual probability; a probability mass function (PMF). It will be clear from the context which is meant. $P(x < a)$ is the probability that x is less than a. An asterisk highlights that a PDF or PMF in unnormalized: P^*. Integrals and sums are implicitly over the full range of the variable unless indicated otherwise, although to avoid ambiguity I indicate the range on sums in some cases (such as when forming numerical averages). Where necessary to distinguish between a quantity x and an estimate thereof, I use a hat – $\hat{x}$ – for the latter.

R code and its screen output are included using the `monospace font`. Shorter pieces of code are provided in the text as part of the discussion. Longer scripts are separated out and are also available online from `http://www.mpia.de/~calj/pbi.html`. A few longer scripts are only available online. You are encouraged to run the examples and to explore by modifying the model parameters and data. The code is provided freely (you can use it in your own work) but without warranty. Most of the plots in this book were produced using the code provided.

No public work can be – or should be – written in isolation. For their useful comments and constructive suggestions I would like to thank Rene Andrae, Brendon Brewer, Morgan Fouesneau, Iskren Georgiev, Nikos Gianniotis, David Hogg, Mike Irwin, Željko Ivezić, Ronald Läsker, Klaus Meisenheimer, Sara Rezaei, Hans-Walter Rix, Jan Rybizki, Luis Sarro, Björn Malte Schäfer, Gregor Seidel, and Ted von Hippel. My special thanks go to my partner Sabina Pauen for supporting (and enduring) me during the writing of this book, and to my parents, Anne Jones and Barrie Jones, from whom I learned so much, and to whom I dedicate this book.

Error and uncertainty are as unavoidable in life as in science. My best efforts notwithstanding, this book inevitably contains errors. Notification of these is appreciated, and a list of corrections will be maintained at `http://www.mpia.de/~calj/pbi.html`.