

Chapter 1 Introduction

Unlikely as it may seem, statistics is currently a sexy subject. Nate Silver's success in out-predicting the political pundits in the last US election drew high profile press coverage across the globe. Statistics may not remain sexy but it will always be useful. It is a key component in the scientific toolbox and one of the main ways we have of describing the natural world and of finding out how it works. In most areas of science, statistics is essential. In some ways this is an odd state of affairs. Mathematical statisticians generally don't require skills from other areas of science in the same way that we scientists need skills from their domain. We have to learn some statistics in addition to our core area of scientific interest. Obviously there are limits to how far most of us can go. This book is intended to introduce some of the most useful applied statistical analyses to researchers, particularly in the life and environmental sciences.

1.1 The approach

My aim is to get across the essence of the statistical ideas necessary to intelligently apply linear models (and some of their extensions) within relevant areas of the life and environmental sciences. I hope it will be of use to students at both undergraduate and post-graduate level and researchers interested in learning more about statistics (or in switching to the software package used here). The approach is therefore not mathematical. I have minimized the number of equations – they are in numerous statistics textbooks and on the internet if you want them – and the statistical concepts and theory are explained in boxes to try and avoid disrupting the flow of the main text. I have also kept citations to a minimum and concentrated them in the text boxes and final chapter (there is no bibliography). Instead, the approach is to learn by doing through the analysis of real data sets. That means using a statistical software package, in this case the R programming language for statistics and graphics (for reasons given below). It also requires data. In fact, most of us only start to take an interest in statistics once we have (or know we soon will have) data. In most science degrees that comes late in the day making teaching introductory statistics more challenging. Students of research degrees (Masters and PhDs) are generally much more motivated to learn statistics. The next best thing to working with our own data is to work with some carefully selected examples from the literature. I have used some data from my own research but I have mainly tried to find small, relevant datasets that have been analysed in an interesting way. Most of them are from the life and environmental sciences (including ecology and evolution). I am very grateful to all of the people who have helped collect these data and to develop the analyses. For convenience I have tried to use data sets already available within the R software (the data sets are listed at the end of the book and described in the relevant chapter).

1.2 The R programming language for statistics and graphics

R is now the principle software for statistics, graphics and programming in many areas of science, both within academia and outside (many large companies use R). There are several reasons for this, including:

- R is a product of the statistical community: it is written by the experts.
- R is free: it costs nothing to download and use, facilitating collaboration.
- R is multiplatform: versions exist for Windows, Mac and Unix.
- R is open-source software that can be easily extended by the R community.
- R is statistical software, a graphics package and a programming language all in one.

1.3 Scope

Statistics can sometimes seem like a huge, bewildering and intimidating collection of tests. To avoid this I have chosen to focus on the linear model framework as the single most useful part of statistics (at least for researchers in the environmental and life sciences). The book starts by introducing several different variations of the basic linear model analysis (analysis of variance, linear regression, analysis of covariance etc.). I then introduce two extensions: generalized linear models (for data with non-normal distributions) and mixed-effects models (for data with multiple levels and hierarchical structure). The book ends by combining these two extensions into generalized linear mixed-effects models. The advantage of following the linear model approach (and these extensions) is that a wide range of different types of data and experimental designs can be analysed with very similar approaches. In particular, all of the analyses covered in this book can be performed in R using only three main functions; one for linear models (the `lm()` function), one for generalized linear models (the `glm()` function) and two for mixed-effects models (the `lmer()` and `glmer()` functions).

1.4 What is not covered

Statistics is a huge subject so lack of space obviously precluded the inclusion of many topics in this book. Many biological approaches like bioinformatics are not covered. I also deliberately left some things out of this book. Due to space, the coverage is limited to linear- and generalized linear models, with nothing on non-linear regression approaches, nor additive models (GAMs). Because of the focus on an estimation-based approach I have not included non-parametric statistics. Experimental design is covered briefly and integrated into the relevant chapters. Information theory and information criteria are briefly introduced but the relatively new and developing area of multi-model inference turned out to be beyond the scope of this book. Bayesian statistics is also completely beyond the scope of the book.

1.5 Approach

There are several different general approaches within statistics (frequentist, Bayesian, information theory etc.) and there are many sub-species within these schools of thought. Most of the methods included of this book are usually described as belonging to 'classical frequentist statistics'. However, this approach, and the probability values that are so widely used within it, has come under increasing criticism. In particular, statisticians usually accuse scientists of focusing far too much on P values and not enough on effect sizes. This is strange as the effect sizes - the estimates and intervals - are directly related to what we measure during our research. I don't know any scientist who studies P values! For that reason I have tried to take an estimation-based approach that focuses on estimates and confidence intervals wherever possible. Styles of analysis vary (and fashions change over time). Because of this I will be frank about some of my personal preferences used in this book. In addition to making wide use of estimates and intervals I have also tried to emphasize the use of graphs for exploring data and presenting results. I have also tried to encourage the use of a priori contrasts (comparisons that were planned in advance) and I avoid the use of corrections for multiple comparisons (and discourage their use in many cases). The most complex approaches in the book are the mixed-effects models. Here I have stuck closely to the approaches advocated by the software writers (and their own books). Finally, at the end of each chapter I try to summarize both the statistical approach and what it has enabled us to learn about the science of each example. It is easy to get lost in statistics but for non-statisticians the analysis should not become an end in its own right, only a method to help advance our science.

1.6 Getting started

To allow a learning-by-doing approach the R code necessary to perform the basic analysis is embedded in the text along with the key output from R (the full R scripts will be available as support material). Some readers may be completely new to R but many will have some familiarity. Rather than start with an introduction to R we will dive straight into the first example of a linear model analysis. However, a brief introduction to R is provided at the end of the book and newcomers to the software will need to start there.