The New Statistics with R

An Introduction for Biologists

ANDY HECTOR

Professor of Ecology
Department of Plant Sciences
University of Oxford

OXFORD
UNIVERSITY PRESS

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 aim of this book

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

The New Statistics with R. Andy Hector. © Andy Hector 2015.
Published 2015 by Oxford University Press.

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 the 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 the teaching of introductory statistics more challenging. Students studying for 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 data sets 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 that are 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 principal 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 (GLMs) (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 classes of function; one for linear models (the lm() function), one for GLMs (the glm() function), and one 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. I also deliberately left some things out. Many biological applications like bioinformatics are not covered. For reasons of space, the coverage is limited to linear models and GLMs, with nothing on non-linear regression approaches nor additive models (generalized additive

2

Comparing Groups: Analysis of Variance

2.1 Introduction

Inbreeding depression is an important issue in the conservation of species that have lost genetic diversity due to a decline in their populations as a result of over-exploitation, habitat fragmentation, or other causes. We begin with some data on this topic collected by Charles Darwin. In *The effects of cross and self-fertilisation in the vegetable kingdom*, published in 1876, Darwin describes how he produced seeds of maize (*Zea mays*) that were fertilized with pollen from the same individual or from a different plant. Pairs of seeds taken from self-fertilized and cross-pollinated plants were then germinated in pots and the height of the young seedlings measured as a surrogate for their evolutionary fitness. Darwin wanted to know whether inbreeding reduced the fitness of the selfed plants. Darwin asked his cousin Francis Galton—a polymath and early statistician famous for 'regression to the mean' (not to mention the silent dog whistle!)—for advice on the analysis. At that time, Galton could only lament that, 'The determination of the variability . . . is a problem of more delicacy than that of determining the means, and I doubt, after making many trials whether it is possible to derive useful conclusions from these few observations. We ought to have measurements of at least fifty plants in each case'. Luckily we can now address this question using any one of several closely related

The New Statistics with R. Andy Hector. © Andy Hector 2015.
Published 2015 by Oxford University Press.

A good place to start is usually by plotting the data in a way that makes sense in terms of our question—in this case by plotting the data divided into the crossed and selfed groups (Fig. 2.1). R has some graphical functions that come as part of the packages that are automatically installed along with the so-called base R installation when you download it from the CRAN website. However, I am going to take the opportunity to also introduce Hadley Wickham's ggplot2 (Grammar of Graphics, version 2) package that is widely used throughout this book. While ggplot2 has an all-singing all-dancing ggplot() function it also contains a handy qplot() function for quickly producing relatively simple plots (and which will take you a surprisingly long way). One advantage of this qplot() function is that its syntax is very similar to that of the base R graphics functions and other widely used R graphics packages such as Deepayan Sarkar's Lattice. Luckily, ggplot2 is supported by a comprehensive website and book so it is easy to expand on the brief introduction and explanations given here. If you do not have the ggplot2 package on your computer you can get it by rerunning the install.packages() function given earlier but substituting ggplot2 in place of SMPracticals. Notice that the qplot() function has a data argument, and one restriction when using ggplot2 is that everything we want to use for

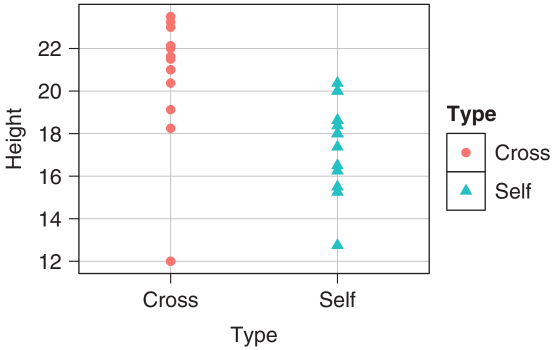

Figure 2.1 The height of Darwin's maize plants (in inches) plotted as a function of the cross- and self-pollinated treatment types. Notice how easy it is with ggplot2 to distinguish treatments with different symbol types, colours (seen as different shades of grey when colour is not available), or both and how a key is automatically generated.

squared, σ^2). The variance is also called the mean square (MS) because it is an average amount of variation: it might be useful to think of it loosely as a per data point average amount of variation. I say 'loosely' because the SS is actually averaged using a mysterious quantity called the degrees of freedom (DF; Box 2.3). The total number of DF is one less than the number of data points (30 – 1 = 29).

However, there's a catch: because of the squaring necessary to calculate the SS this estimate of the variability is not on the same scale as the original

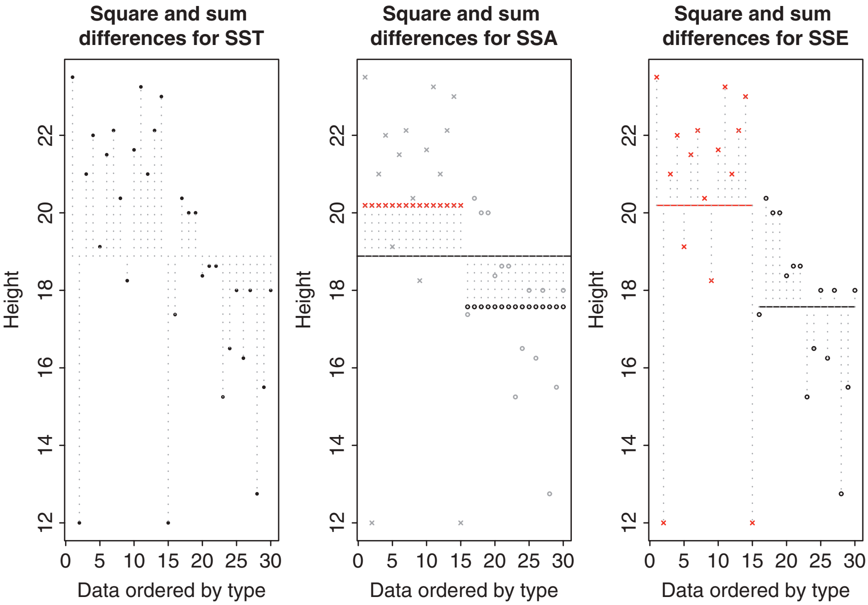

Figure 2.2 How to calculate the total (SST), treatment (SSA), and error (SSE) sums of squares. In each panel the vertical lines measure the differences that are then squared and summed. The SST (left) is calculated by measuring the differences from each data point to some reference point—the overall mean is the most intuitive one for teaching purposes as shown by the horizontal line (although for technical reasons it is generally not the one used by statistical software!). The differences for the SSA (middle) are between the treatment-level means (the horizontal lines of 'fitted values' shown by the crosses and circles) and the grand mean. The differences for the SSE (right) are between the observed data values and the treatment-level means.

form of the data set that classifies the columns into the identity variables that define its structure and measure variables—in this case just our single response variable:

```
> mDarwin <- melt(darwin, id.vars= c("pot", "pair", "type"),
> measure.vars="height") # not shown
```

Then we can cast a wide version of the data set that specifies pot and pair to be the rows (given on the left-hand side of the tilde, ~) and forming a column of heights for each pollination type (as specified on the right-hand side of the tilde—note that because we only measured one response the column labelled 'variable' has only this one repeated entry all the way down):

```
> darwide <- cast(mDarwin, pot + pair ~ variable + type)
> head(darwide)
```

```
   pot  pair  height_Cross  height_Self
1    I     1        23.500       17.375
2    I     2        12.000       20.375
3    I     3        21.000       20.000
4   II     4        22.000       20.000
5   II     5        19.125       18.375
6   II     6        21.500       18.625
```

The head() function allows us to just output the first few rows of the reshaped data frame. By substituting `darwide$height.cross` and `darwide$height.self` into the functions given earlier for the mean and SD of the long version of the data we can see that the crossed plants have a height of 20.2 while the selfed plants have a lower mean height of 17.6, a shortfall of about 2.6 inches. The question is: how confident are we that this difference reflects negative effects of selfing? To judge this we have to assess the signal (the difference between treatment means) relative to the noise (the level of variation within the samples). ANOVA uses variance to quantify the signal and noise, but to get a first impression of the level of variability on the same scale as the means we can apply the sd() function to the wide data frame to get values for the SDs of the crossed and selfed samples of 3.6 and 2.1 inches, respectively (see Section 2.2).