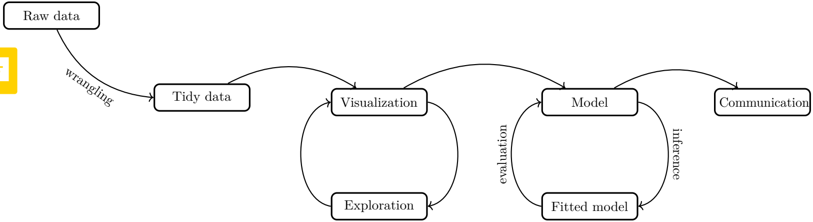
Chapter 1: Data Analysis and Data Science

Mark Andrews
Psychology Department, Nottingham Trent University

✉ mark.andrews@ntu.ac.uk

Introduction

This book is about statistical data analysis of real world data using modern tools. It is aimed at those who are currently engaged in, or planning to be engaged in, analysis of statistical data of the kind that might arise at or beyond PhD level scientific research, especially in the social sciences. The data in these fields is complex. There are many variables and complex relationship between them. Analyzing this data almost always requires data wrangling, exploration, and visualization. Above all, it involves modelling the data using flexible probabilistic models. These models are then used to reason and make predictions about the scientific phenomenon being studied. This book aims to address all of these topics. The term we use for these topics and their corresponding methods and tools is *data science*.



What is data science?

Even if we accept the nature and the value of the data analysis workflow that we've just outlined, it is reasonable to ask whether it should properly be called data science. Is this not just using a new word, even a buzzword, in place of much more established terms like statistics or statistical data analysis? We are using the term data science rather than statistics per se or some variant thereof because data analysis as we've outlined it arguably goes beyond the usual focus of statistics, at least as it is traditionally understood. Mathematical statistics as a scientific or mathematical discipline has focused largely on the statistical modelling component of the program we outlined above. As we've hopefully made clear, this component is of profound importance, and in fact we would argue that it is the single most important part and even ultimate goal of data analysis. Nonetheless, in practice, data wrangling alone occupies far more of our time and effort in any analysis, and exploration and visualization should be seen as necessary precursors to, and even continuous with, the statistical modelling itself. Likewise, traditional statistics often marginalizes the practical matter of computing tools. In statistics textbooks, even excellent ones, for example, code examples may not be provided for all analyses, and the code may not be integrated tightly with the coverage of the statistical methods. In this sense, traditional statistics does not

Why R, not Python?

We have stated repeatedly that computational methods and tools are vital for doing data science. In this book, the computing language and environment that we use is *The R Project for Statistical Computing*, simply known as R. More specifically, we use the modern incarnation of R that is based on the so-called *tidyverse*. In Chapter 2, we provide a proper introduction to R. Here, we wish to just outline why R is our choice of language and environment, what the alternatives are, and what this entails in terms of the our conception of what data science is and how it is practiced.

Given our conception of the data science workflow that we outlined in Figure 1, R is an inevitable choice. We believe that R is simply the best option to perform all the major components that we outline there. For example, for the data wrangling component, which can be extremely labourious, R packages that are part of the tidyverse such as `readr`, `dplyr`, `tidyr` and so on, which we cover in Chapter 3, makes data wrangling fast and efficient and even pleasurable. For data visualization, the `ggplot2` package provides us with essentially a high level and expressive language for data visualization. For the statistical modelling loop, which we cover in all the chapters of Part II of this book, R provides a huge treasure trove of packages for virtually every

Who is this book for?

As mentioned at the start of this chapter, the prototypical audience at whom this book is aimed are those engaged in data analysis in scientific research, specifically research at or beyond PhD level. In scientific research, statistics obviously plays a vital role, and specifically this is based on using data to build and interpret statistical or probabilistic models of the scientific phenomenon being studied. This book is heavily focused on this particular kind of statistical data analysis. As we've mentioned, in data science as it is practiced in industry and business, often the other "culture" of statistics (see Breiman 2001), namely predictive analytics and algorithms, is the dominant one, and so this book is not ideal for those whose primary data science interests are of that kind.

We've explicitly stated that this book is intended for those doing research in the social sciences, but this also requires some explanation. The explicit targeting of the social sciences is largely just to keep some focus and limits to the sets of examples that are used throughout the book. However, beyond the example data sets that are used, there is little about of this content that is of relevance to only those doing research in social science disciplines. All the content on data wrangling, exploration and visualization, statistical modelling, etc., is hopefully

The style and structure of this book

Apart from this brief introductory chapter, all the remainder of the book is a blend of expository text, R code, mathematical equations, diagrams and R based plots. It is intended that people will read this book while using R to execute all the code examples and so produce all the results that are presented either as R output or as figures. Of course, if readers prefer to read first and the run the code later, perhaps on a second reading, that is entirely a matter of preference. However, all the code that we present throughout this book is ready to run, and does not require anything other than the R packages that are explicitly mentioned in the code and the all the data sets that are being used, which are all available in the website that accompanies this book.

The book is divided into three parts. Part I is all about the parts of the data science workflow shown in Figure 1 except for the statistical modelling loop part. Thus, in Part I we provide a comprehensive general introduction to R, a chapter on data wrangling using `dplyr`, `tidyr` etc., a chapter on data visualization, and another on data exploration. In another chapter, we go into more detail about programming in R, and provide a chapter on doing reproducible data analysis using tools like RMarkdown and Git. Part II of the book, which is the largest part, is all about the statistical modelling loop part

Reference

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).” *Statistical Science* 16 (3): 199–231.

Cleveland, William S. 2001. “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.” *International Statistical Review* 69 (1): 21–26.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver; Boyd.

Tukey, John W. 1962. “The Future of Data Analysis.” *Annals of Mathematical Statistics* 33 (1): 1–67.