Analyzing Linguistic Data

A Practical Introduction to Statistics Using R

Statistical analysis is a useful skill for linguists and psycholinguists, allowing them to understand the quantitative structure of their data. This textbook provides a straightforward introduction to the statistical analysis of language data. Designed for linguists with a non-mathematical background, it clearly introduces the basic principles and methods of statistical analysis, using R, the leading computational statistics programming environment. The reader is guided step-by-step through a range of real data sets, allowing them to analyze phonetic data, construct phylogenetic trees, quantify register variation in corpus linguistics, and analyze experimental data using state-of-the-art models. The visualization of data plays a key role, both in the early stages of data exploration and later on when the reader is encouraged to criticize initial models fitted to the data. Containing over 40 exercises with model answers, this book will be welcomed by all linguists wishing to learn more about working with and presenting quantitative data.

The program R is available at http://cran.at.r-project.org/. The data sets and ancillary functions discussed in this book have been brought together in the language R package, which is available at the same URL.

R. H. BAAYEN is Professor of Quantitative Linguistics at the University of Alberta, Edmonton. He is author of *Word Frequency Distributions* (2001), co-editor of *Morphological Structure in Language Processing* (2003), and has published widely in linguistics and psycholinguistics journals.

Analyzing Linguistic Data

A Practical Introduction to Statistics Using R

R. H. BAAYEN

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521709187

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-88259-0 hardback
ISBN 978-0-521-70918-7 paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this book, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To Jorn, Corine, Thera, and Tineke

Contents

Cambridge University Press
978-0-521-88259-0 - Analyzing Linguistic Data: A Practical Introduction to Statistics Using R
R. H. Baayen
Frontmatter
More information

Preface

This book provides an introduction to the statistical analysis of quantitative data for researchers studying aspects of language and language processing. The statistical analysis of quantitative data is often seen as an onerous task that we would rather leave to others. Statistical packages tend to be used as a kind of oracle, from which you elicit a verdict as to whether you have one or more significant effects in your data. In order to elicit a response from the oracle, you have to click your way through cascades of menus. After a magic button press, voluminous output tends to be produced that hides the *p*-values, the ultimate goal of the statistical pilgrimage, among lots of other numbers that are completely meaningless to the user, as befits a true oracle.

The approach to data analysis to which this book provides a guide is fundamentally different in several ways. First of all, we will make use of a radically different tool for doing statistics, the interactive programming environment known as `R`. `R` is an open source implementation of the (object-oriented) S language for statistical analysis originally developed at Bell Laboratories. It is the platform par excellence for research and development in computational statistics. It can be downloaded from the COMPREHENSIVE R ARCHIVE NETWORK (CRAN) at `http://cran.r-project.org` or one of the many mirror sites. Learning to work with `R` is in many ways similar to learning a new language. Once you have mastered its grammar, and once you have acquired some basic vocabulary, you will also have begun to acquire a new way of thinking about data analysis that is essential for understanding the structure in your data. The design of `R` is especially elegant in that it has a consistent uniform syntax for specifying statistical models, no matter which type of model is being fitted.

What is essential about working with `R`, and this brings us to the second difference in our approach, is that we will depend heavily on *visualization*. `R` has outstanding graphical facilities, which generally provide far more insight into the data than long lists of statistics that depend on often questionable simplifying assumptions. That is, this book provides an introduction to *exploratory data analysis*. Moreover, we will work *incrementally* and *interactively*. The process of understanding the structure in your data is almost always an iterative process involving graphical inspection, model building, more graphical inspection, updating and adjusting the model, etc. The flexibility of `R` is crucial for making this iterative process of coming to grips with your data both easy and in fact quite enjoyable.

A third, at first sight heretical aspect of this book is that I have avoided all formal mathematics. The focus of this introduction is on explaining the key concepts and on providing guidelines for the proper use of statistical techniques. A useful metaphor is learning to drive a car. In order to drive a car, you need to know the position and function of tools such as the steering wheel and the brake pedal. You also need to know that you should not drive with the handbrake on. And you need to know the traffic rules. Without these three kinds of knowledge, driving a car is extremely dangerous. What you do not need to know is how to construct a combustion engine, or how to drill for oil and refine it so that you can use it to fuel that combustion engine. The aim of this book is to provide you with a driving licence for exploratory data analysis. There is one caveat here. To stretch the metaphor to its limit: with R, you are receiving driving lessons in an all-powerful car, a combination of a racing car, a lorry, a family car, and a limousine. Consequently, you have to be a responsible driver, which means that you will find that you will need many additional driving lessons beyond those offered in this book. Moreover, it never hurts to consult professional drivers—statisticians with a solid background in mathematical statistics who know the ins and outs of the tools and techniques, and their advantages and disadvantages. Other introductions that you may want to consider are Dalgaard (2002), Verzani (2005), and Crawley (2002). The present book is written for readers with little or no programming experience. Readers interested in the R language itself should consult Becker *et al.* (1988) and Venables and Ripley (2002).

The approach I have taken in this course is to work with real data sets rather than with small artificial examples. Real data are often messy, and it is important to know how to proceed when the data display all kinds of problems that standard introductory textbooks hardly ever mention. Unless stated otherwise, data sets discussed in this book are available in the `languageR` package, which is available at the CRAN archives. You are encouraged to work through the examples with the actual data, to get a feeling for what the data look like and how to work with R's functions. To save typing, you can copy and paste the R code of the examples in this book into the R console (see the file `examples.txt` in `languageR`'s `scripts` directory). The `languageR` package also makes available a series of functions. These convenience functions, some of which are still being developed, bear the extension `.fnc` to distinguish them from the well-tested functions of R and its standard packages.

An important reason for using R is that it is a carefully designed programming environment that allows you, in a very flexible way, to write your own code, or modify existing code, to tailor R to your specific needs. To see why this is useful, consider a researcher studying similarities in meaning and form for a large number of words. Suppose that a separate model needs to be fitted for each of 1000 words to the data of the other 999 words. If you are used to thinking about statistical questions as paths through cascaded menus, you will discard such an analysis as impractical almost immediately. When you work in R, you simply write the code for one word, and then cycle it through on all other words. Researchers are often unnecessarily limited in the questions they explore because they are thinking in a menu-driven language instead of in an interactive programming language like R. This is an area where language determines thought.

If you are new to working with a programming language, you will find that you will have to get used to getting your commands for R exactly right. R offers command line editing facilities, and you can also page through earlier commands with the up and down arrows of your keyboard. It is often useful to open a simple text editor (emacs, gvim, notepad), to prepare your commands in, and to copy and paste these commands into the R window, especially as more complex commands tend to be used more than once, and it is often much easier to make copies in the editor and modify these, than to try to edit multiple-line commands in the R window itself. Output from R that is worth remembering can be pasted back into the editor, which in this way comes to retain a detailed history both of your commands and of the relevant results. You might think that using a graphical user interface would work more quickly, in which case you may want to consider using the commercial software S-PLUS, which offers such an interface. However, as pointed out by Crawley (2002), "If you enjoy wasting time, you can pull down the menus and click in the dialog boxes to your heart's content. However, this takes about 5 to 10 times as long as writing in the command line. Life is short. Use the command line" (p. 11).

There are several ways in which you can use this book. If you use this book as an introduction to statistics, it is important to work through the examples, not only by reading them through, but by trying them out in R. Each chapter also comes with a set of problems, with worked-out solutions in Appendix A. If you use this book to learn how to apply in R particular techniques that you are already familiar with, then the quickest way to proceed is to study the structure of the relevant data files used to illustrate the technique. Once you have understood how the data are to be organized, you can load the data into R and try out the example. And once you have got this working, it should not be difficult to try out the same technique on your own data.

This book is organized as follows: Chapter 1 is an introduction to the basics of R. It explains how to load data into R, and how to work with data from the command line. Chapter 2 introduces a number of important visualization techniques. Chapter 3 discusses probability distributions, and Chapter 4 provides a guide to standard statistical tests for single random variables as well as for two random variables. Chapter 5 discusses methods for clustering and classification. Chapter 6 discusses regression modeling strategies, and Chapter 7 introduces mixed-effects models, the models required for analyzing data sets with nested or crossed repeated measures.

I am indebted to Carmel O'Shannessy for allowing me to use her data on Warlpiri, to Kors Perdijk for sharing his work on the reading skills of young children, to Joan Bresnan for her data on the dative alternation in English, to Maria Spassova for her data on Spanish authorial hands, to Karen Keune for her materials on social and geographical variation in the Netherlands and Flanders, to Laura de

Vaan for her experiments on Dutch derivational neologisms, to Mirjam Ernestus for her phonological data on final devoicing, to Wieke Tabak for her data on etymological age, to Jen Hay for the rating data sets, and to Michael Dunn for his data on the phylogenetic classification of Papuan and Oceanic languages. I am also grateful to Adrian Stenton for his careful copy-editing of the manuscript. Many students and colleagues have helped me with their comments and suggestions for improvement. I would like to mention by name Joan Bresnan, Mirjam Ernestus, Jen Hay, Reinhold Kliegl, Victor Kuperman, Petar Milin, Ingo Plag, Hedderik van Rijn, Stuart Robinson, Eva Smolka, and Fiona Tweedie. I am especially indebted to Douglas Bates for his detailed comments on Chapter 7, his advice for improving the `languageR` package, his help with the code for temporary ancillary functions for mixed-effects modeling, and the insights offered on mixed-effects modeling. In fact, I would like to thank Doug here for all the work he has put into developing the `lme4` package, which I believe is the most exciting tool discussed in this book for analyzing linguistic experimental data. Last but not least, I am grateful to Tineke for her friendship and support.

In this book, small capitals denote key concepts and technical terms. Typewriter font is used for R code and R objects. Linguistic examples are typeset with italics, as are statistical symbols.