Journal of Research Design and Statistics in Linguistics and Communication Science

JRDS (PRINT) ISSN 2052-417X
JRDS (ONLINE) ISSN 2052-4188

Review

How to do Linguistics with R – Data Exploration and Statistical Analysis

Natalia Levshina (2015)

Reviewed by Andreea S. Calude

Natalia Levshina's (NL) *How to do Linguistics with R – Data Exploration and Statistical Analysis* appears to be (at least to my knowledge) the fourth book dedicated specifically to linguistics research employing statistical analyses in R (R Development Core Team, 2009). Following Baayen's 2008 *Analysing Linguistic Data: A practical Introduction to Statistics Using R*, and Gries' two 2009 books, *Statistics for Linguists with R: A Practical Introduction* and *Quantitative Corpus Linguistics with R: A Practical Introduction*, the recent addition does indeed fill a need in the linguistics community for more statistical information. Within an ever-growing landscape of corpus linguistics studies, large data sets, increased computational power and increased computer storage space, many researchers are finding that a quantitative approach can fruitfully be used to pursue novel and innovative questions.

In comparison to its predecessors, the title of the current book distinguishes itself by not promising to be 'practical' or an 'introduction'. As it turns out, it is 'practical' in its use of real data sets, much like Baayen and Gries, as well as the fact that the methods discussed come with advice regarding both usability and pitfalls. However, an 'introduction' it is not. The book is best suited for researchers who already have some basic working knowledge of statistics and perhaps even of R (or at the very least for those who are armed with sufficient confidence to approach the subject). Its goal is to '*provide a linguist with a statistical toolkit*' (jacket cover). In other words, unlike the Baayen or

Affiliation

University of Waikato, New Zealand.
email: acalude@gmail.com

JRDS VOL 3.2 2016 268–272
https://doi.org/10.1558/jrds.30361

EQUINOX
equinoxonline

Gries texts, this is not a course in statistics per se (although Levshina refers to it as a 'textbook', page 5), but an array of different methods available, each with a sufficiently comprehensive description to offer a place to start and a direction for pursuing further information about each technique discussed. At the same time, the book is kept suitably brief so as not to overwhelm. Owing to its encyclopaedic function, the book does not assume the reader will consume the book chapters in the order presented, and I find this to be a great strength.

One important aspect to discuss here is the choice of statistical software. The reason why linguists are increasingly turning to R in favour of packages like SPSS or JMP is that R is powerful, freely available, and it can run on virtually any machine (for example, gone are the days when Mac users had to contend with Mac-only software). One (initial but quickly overcome) downside of R is that loading data and running commands in R can be a daunting task to begin with, since it is not a WYSIWYG (*what-you-see-is-what-you-get*) software. However, with the RStudio[1] interface (RStudio Team 2015), described by NL (pp. 32), this potential shortcoming can be minimized, if not altogether eliminated.

The book is organized into 20 chapters. The first three chapters (pp. 1–68) function as a whirlwind introduction, taking readers through a rapid tour of what statistics can and cannot do, how to formulate research hypotheses, what major types of statistic tests are available (parametric and non-parametric tests), what R is and how to install it, how to install the *Rling* package used in the book (which contains data sets and a few functions created by Levshina), different types of variables, basic statistical measures (mean, median, mode) and (basic) visualizations of these.

There are two chapters which focus exclusively on visualization of data, namely chapter 4 (pp 69–86) which deals with the visualization of qualitative variables, and chapter 20 (pp. 387–394) which looks at motion charts as a means for visualizing language change. Data plots and other graphic displays are included as appropriate through the rest of the book, but these do not take centre stage in the remaining chapters. The R code used to draw the various plots is given in both, standard plotting R functions, as well as the widely used *ggplot2* commands (Wickham and Winston, 2009).

Alongside chapters 4 and 20, chapters 5–19 (pp. 87–387) form the main contents of the book. They detail a range of statistical tests and analyses available for different types of data and used to answer a wide selection of research questions, including *t*-test, Wilcoxon and Mann-Whitney test, linear regression, analysis of variance (ANOVA), association measures, probabilistic multifactorial analysis, logistic regression, conditional inference trees, random forests, behavioural profiles, distance measures, cluster analysis, semantic vector spaces, multidimensional scaling, principal component analysis, factor analy-

sis, and correspondence analysis. Each chapter is meant as a stand-alone section, although good cross-reference links are provided to direct the reader as necessary. The focus of each chapter is on a specific family of related methods, with the more well-known or more commonly used being presented first and in most detail. The method is described with reference to the research hypotheses associated with it, the type of data required to answer these (a real data set is assigned to each chapter for the reader to practise on, downloadable through *Rling*), details of how to report the results from the statistical tests performed, relevant ways of visualizing the data where appropriate, and further reading for those requiring a more in-depth treatment of the topic. In particular, the book's value is greatly increased by two aspects included in each chapter: (1) details of how to report the findings of the statistics test; this being an aspect often forgotten or ignored in other texts, and (2) details of particular linguistics research paper(s) which implement the statistical test described.

The book has a companion site (https://benjamins.com/sites/z.195/) which includes exercises and questions relevant to its chapters (no answers are provided), R-code used throughout the book (available to copy-and-paste), and the *Rling* package (available to download). It also includes a reference appendix of frequently used and useful R commands (pp. 397–408) and of main plotting functions (pp. 409–424). These are all helpful resources.

NL makes good use of clear language and easy to follow examples in her book. I found the graphical display of the contents visually appealing and pleasant to engage with. Only one instance of confusion comes to mind (in chapter 10, page 227, the total number of verbs in the ditransitive construction found in the corpus does not match the total number presented in the data file associated with this example set, and the same goes for the total number of verb constructions – it turns out that this is because the data file included in *Rling* only includes a sub-portion of the total verbs in the corpus), which was easily clarified by an email exchange with the author. Overall, the book stands out in its readability.

A further strength of NL's book is that it provides a wide breadth of statistical techniques, many of which are not documented in other texts intended specifically for linguistics. This is particularly useful, as language data is notoriously diverse in nature, most often violates parametric test assumptions (such as normal distributions), and linguists tend to find themselves in the rather awkward position of having to develop knowledge and experience with a wide array of statistical tests without knowing quite where or how to begin. Therefore, this book can serve as an important point of orientation for the linguistics researcher in the statistics landscape.

Second, *How to do Linguistics with R – Data Exploration and Statistical Analysis* gives good practical tips associated with the various tests described

(for example, short explanations, such as why one might choose a parametric test in favour of a non-parametric one given the cumbersome assumptions that parametric tests require). As mentioned above, the book is best used as a starting point in order to guide one's reading towards the relevant statistical tests for the particular data at hand. Because of the brevity of each chapter, it is unlikely that the book will equip the reader with the full details of what is involved in each test, and caution might need to be exercised in this respect. For example, in chapter 11, the discussion of the collexeme analysis walks the reader through the necessary steps and commands to perform the collexeme analysis of adjectives modified by *quite*, but the explanation does not provide a full mathematical description (akin to say Gries, 2009a) of the effect of the logarithmic function on *p*-values (beyond stating that it would improve their interpretability). This need not be understood as a criticism of the book but more as a natural limitation, since given the wide breadth of its coverage, it is expected that many details will be left out (otherwise, we might be looking at a book comparable to Crawley's 2015 manual).

The book represents a worthy contribution to the growing literature encompassing statistical methodology texts aimed at the linguistics researcher, and fills a needed gap as a compendium of diverse techniques available for investigating language data. While in my opinion, the book is not suitable for the complete newcomer to statistics, I would recommend it as a port-of-call for those have used R before but are still in need a bit of help finding their way through the web of methods available for analysis of quantitative data. Brian Joseph's comment aptly captures the high demand and relevance of such a book in the increasingly empirically-driven discipline of linguistics: '*Linguistics has always had a numerical and mathematical side […], but the use of quantitative methods, and, relatedly, formalizations and modeling, seems to be ever on the increase; rare is the paper that does not report on some statistical analysis of relevant data or offer some model of the problem at hand*' (Joseph, 2008: 687 cited in Gries, 2013: 361).

Note

1. I personally do not use RStudio but I understand that the code given in the book is not as user-friendly in RStudio because the syntax highlighting is lost.

References

Baayen, H. (2008). *Analysing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Crawley, M. (2015). *Statistics: An Introduction using R* [2nd edition]. Chichester: Wiley & Sons.

Gries, S. T. (2009a). *Statistics for Linguists with R: A Practical Introduction*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110216042

Gries, S. T. (2009b). *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge. https://doi.org/10.1515/9783110216042

Gries, S. T. (2013). Elementary statistical testing with R. In M. Krug and J. Schlüter (Eds), *Research Methods in Language Variation and Change*, 361–381. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511792519.024

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. http:www.R-project.org (1 April 2016).

RStudio Team (2015). *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Levshina, N. (2015). *Rling*. Data sets and functions with *How to do Linguistics with R – Data Exploration and Statistical Analysis*. Available to download from web address: https://benjamins.com/sites/z.195/content/package.html.

Wickham, H. and Winston, C. (2009) *ggplot2: An Implementation of the Grammar of Graphics*. http://docs.ggplot2.org/ [accessed 14 March 2016].