Statistics for linguistics with R: A practical introduction (review)

Steven J. Clancy

Language, Volume 88, Number 2, June 2012, pp. 426-429 (Article)

Published by Linguistic Society of America
DOI: 10.1353/lan.2012.0032

Statistics for linguistics with R: A practical introduction. By STEFAN TH. GRIES. (Trends in linguistics: Studies and monographs 208.) Berlin: Mouton de Gruyter, 2009. Pp. x, 335. ISBN 9783110205657. $49.95.

Reviewed by STEVEN J. CLANCY, *University of Chicago*

This is a fortunate time to be a linguist seeking the right tools for taking your research in a more statistical and quantitative direction. The past few years have witnessed the arrival of several excellent books, not only designed as introductions to statistics or data analysis in general, but specifically tailored to the interests and needs of linguists (Baayen 2008, Gries 2009 and the volume under review, Johnson 2008). These books share a promotion of R, a multiplatform, open-source software package that functions as a programming language, a statistical environment, a graphics package, and a superb all-around tool for data processing, storage, and manipulation (www.r-project.org). Anyone interested in learning more about R, corpus linguistics, and statistical methods in linguistic research will be well served by these books and will find the two books by Stefan Th. Gries to be indispensable learning tools. G's *Quantitative corpus linguistics with R* (2009) and the current volume, *Statistics for linguistics with R*, provide excellent introductions to statistics and quantitative methods in linguistics, to the R environment, to corpus linguistics, and to the general ways of thinking and of formulating hypotheses necessary for quantitative linguistic research.

This volume is intended as an introductory textbook in using statistics in linguistic research and to the kinds of questions one asks in developing, formulating, and testing hypotheses. The book consists of five content chapters and an epilogue, and is accompanied by a website and a newsgroup along with downloadable code, exercises, data, and an answer key. Beginning with 'Some fundamentals of empirical research' (Ch. 1), the author introduces the reader to the collection and analysis of experimental and corpus data. A particularly strong point of the book is that G provides numerous models and templates for conducting the kind of research he advocates, much as he does in Gries 2009 and his article on more rigorous methods in corpus linguistics (Gries 2006).

G then covers the 'Fundamentals of R' (Ch. 2), providing the reader with guidance in setting up R and starting out with the first few functions and lines of code. This introduction to R is thorough enough to give the reader confidence and comfort with the R environment and managing data structures. R is freely available for Windows, Mac, and Linux platforms, but the discussion here is somewhat biased toward Windows users. Although there is a significant learning curve in becoming a proficient user of R, G's and Baayen's books are more than sufficient to teach you what you need to know in order to do serious linguistic research with R. As a programming language, R is so useful and powerful that one might even forego learning other programming languages such as Python, at least for a time. G is extremely encouraging throughout. His book 'aims to help you do scientific quantitative research' (2), and it succeeds in reaching this goal. Along the way, G helps readers deal with frustration (54) and keeps up their morale with encouraging comments.

As an introduction to using R itself, Gries 2009 is more comprehensive, but the current book has the task of introducing the reader to statistical techniques and how to employ them. In Chs. 3–5, G begins a whirlwind tour of statistical tests and methods that linguists will find useful. Most statistical tests are carried out on a variety of actual linguistic datasets included with the book at the companion website, making the discussion and plots that much more relevant and tailored to the needs of linguists. This approach to thinking about how the various methods might be used in your own research, rather than merely absorbing material about statistics from a general introductory text, appreciably increases the value of the book. G presents a few equations throughout the book, but the level of math required is no higher than what one would commonly know from high school and basic college math courses. G shows how to calculate the statistics step-by-step for simpler techniques, then shows how to do things with single R functions. For more complex procedures, G walks the reader through ready-made functions and scripts, providing recommendations for further reading on the mathematics involved for the curious.

The reader will learn about different types of variables (independent, dependent; nominal, binary, categorical, ordinal, ratio), about formulating falsifiable and testable hypotheses, and about the alternative hypothesis (H_1) and the null hypothesis (H_0) and when to accept the one and reject the other. Readers will learn about the statistical use of the word *significant*, probability of error, normal distribution, and one-tailed and two-tailed tests, among other topics and terms. G provides a fair amount of guidance in analytical reasoning, operationalizing variables, and how to report one's results in an article. Graduate students and statistical neophytes will find these models quite useful and motivating. Throughout, readers will learn how to prepare and import their data into R and carry out these statistical calculations, including ample practice with the graphing capabilities in R.

'Descriptive statistics' (Ch. 3) discusses univariate and bivariate statistics and covers scatterplots, line plots, pie charts, bar plots, pareto-charts, histograms, mosaic plots, spineplots, and interaction plots. G provides instruction in measures of central tendency (mode, median, mean), measures of dispersion, deviation, error, *z*-scores, confidence intervals, and linear regression. The techniques in 'Analytical statistics' (Ch. 4) allow the user to test for significance and to compare samples. In sections on distribution, dispersion, means, and correlation, readers learn the how and why behind the statistics they may previously have seen in papers but not fully understood. G covers quantifying correlations with Cramer's *V* values, testing for normality with the Shapiro-Wilk test, and the use of the chi-square test to compare observed and expected distributions. Subsections present a different type of data or question to be answered and methodically present how to approach the problems, formulate hypotheses, carry out the necessary tests and calculations, and report the results. Users learn to compare means with *t*-tests and other measures, calculate probabilities of error, compute effect sizes, and measure correlations.

Whereas Chs. 3 and 4 detailed 'how maximally one independent variable is correlated with the behavior of one dependent variable' (238), G then presents the challenges and tools for 'Selected multifactorial methods' (Ch. 5). These techniques are among the most useful for linguists since the phenomena we explore are so often layered with meaning, and usage is affected by many factors for any given construction. G covers configural frequency analysis, hierarchical configural frequency analysis, multiple regression analysis, analysis of variance (ANOVA), and hierarchical agglomerative cluster analysis. In a book packed full of terms, acronyms, equations, and complex procedural descriptions, G almost never takes his eye off the ball. I only noticed two instances where something was not explained sufficiently. One was a missing explanation for the acronym AIC (Akaike information criterion, p. 239 and used throughout Ch. 5), and the other is a presentation of a pairwise scatterplot (256) without any explanation of how to read these complex, densely packed diagrams. The former slip is quite minor, but the latter should be remedied in a future edition (cf. Baayen 2008:36–37 on scatterplot matrices for a guide to these plots). This chapter is perhaps the most complex chapter of the book and one that rewards careful reading and implementation of the provided code. I second G's recommendation early on in the book that you must work your way through the examples, rather than simply reading through the discussion in the book, even if you are already an experienced user of R. For instance, my own understanding of how to apply multiple regression analysis was hampered until I thoroughly worked through the provided code and made my own decisions at each step. Although the book is not overly long, the topics are often covered with a density that requires significant time for both reading and working with R.

G notes that the 'number of studies utilizing quantitative methods has been increasing (in all linguistic sub-disciplines)' and that the 'field is experiencing a paradigm shift towards more empirical methods' (4). Despite this growth, he acknowledges that one sometimes encounters reluctance and resistance among linguists to these methods. In order to dispel the potentially revolutionary character of quantitative approaches, he notes reassuringly that 'in practice quantitative and qualitative methods go hand in hand: qualitative considerations PRECEDE and FOLLOW the results of quantitative methods' (4, emphasis mine). G notes that linguists are not done away with by such methods, but rather our input is needed at all stages, yet we are all greatly assisted by the objective guides such methods provide (cf. p. 308).

One of the great strengths of this book is that researchers at any level (undergraduate and graduate students, post-docs, and experienced faculty) can use it to learn about the techniques and try them out using meaningful linguistic data, while figuring out how to apply these topics to their own problems of interest. They will then be ready to use the models presented in order to learn how to talk about and report the statistics they compute. As a first step, one needs to ask and answer such questions as: What are the tests and methods at my disposal and when do I use them? What other methods can I use to explore my data, reveal structure in it, and form hypotheses? Such statistical tools become even more powerful when you know more about them and how to choose from among them, and this book will help you get to that point. G stresses the importance of understanding your data, and thinking about how you will operationalize your variables, and the importance of visual inspection of various plots even before carrying out the necessary statistical tests. For instance, users learn to create and interpret boxplots, including work with whiskers, notches, means, and outliers, and learn how visual inspection often reveals what is later confirmed in the statistical calculations.

Among other comments on the book, I would note that all code and output in R is quite clearly presented in a fixed-width typeface in gray boxes. Any function names or parameters are referenced in this same typeface in the body of the text. The author has chosen to use a small bullet character (•) to indicate spaces and a paragraph sign (¶) to indicate where the user should enter carriage returns when typing in commands. This convention does effectively resolve any possible ambiguities in the code, but makes the code a bit cluttered and less readable by users when comparing the book's code to what they are seeing on their consoles. The code is generally quite accurate as well. Two code errors I detected in working through the book (209, 264) were both already noted in the errata file on the companion website.[1] Until a new version of the book is in print, it would behoove the reader to take a look at this errata page in order to avoid any possible confusion. In terms of general presentation, one encounters some typos, small errors, and awkward turns of phrase (perhaps due to translation from the original German version). Although the book would benefit from another round of close proofreading and editing for a future edition, none of these issues mar what is otherwise a model of clear instruction and description of quite complex material.

The print edition of the book is easy to find at amazon.com and elsewhere, but is difficult to find at the De Gruyter website. The publisher made an ebook version available for this review, but when I searched for this version of the book on the internet, I could only find it listed at De Gruyter Mouton for $1,372![2] This is unfortunate, because an ebook with a full-text search would make the book very useful as a reference tool, allowing one to search for code and techniques that may not be found in either the table of contents or the index, which is an index only of functions in R and not of general topics covered in the book.

Individual users will profit from reading G's book and instructors will find this book (with plentiful data samples, 'think breaks', and exercises) useful as a course textbook. There is also much to be gained by using G's books in conjunction with other books on R. Particularly for those interested in data-visualization methods, Baayen 2008 and Yau 2011 have many interesting examples. I would also recommend that those interested seek out an opportunity to attend a workshop on empirical methods (upcoming offerings include G's own 'boot camps' at UC Santa Barbara in July–August 2012 or the 'Empirical methods in cognitive linguistics' series with its next installment in August 2012 at Case Western University). Hands-on instruction and practice with R and with these methods is critical for developing confidence in using R and for becoming comfortable with these techniques and their application. With his two books and numerous articles on these topics and frequent participation in workshops, boot camps, and master classes on empirical methods, G has made a major contribution toward the promotion of empirical methods in linguistics and toward the training of a new generation of linguists.

[1] The errata page may be found at http://www.linguistics.ucsb.edu/faculty/stgries/research/sflwr/e_errata.txt.

[2] http://www.degruyter.com/view/product/38552?rskey=2umYBT&result=1&q=stefan%20th.%20gries

REFERENCES

BAAYEN, R. HARALD. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

GRIES, STEFAN TH. 2006. Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54.2.191–202.

GRIES, STEFAN TH. 2009. *Quantitative corpus linguistics with R: A practical introduction*. London: Routledge.

JOHNSON, KEITH. 2008. *Quantitative methods in linguistics*. Oxford: Blackwell.

YAU, NATHAN. 2011. *Visualize this: The flowing data guide to design, visualization, and statistics*. Indianapolis: Wiley.

University of Chicago
Department of Slavic Languages & Literatures
1130 E. 59th St.
Chicago, IL 60637
[sclancy@uchicago.edu]

Anthropological papers of the University of Alaska: The Dene-Yeniseian connection. Ed. by JAMES KARI and BEN A. POTTER. Fairbanks: University of Alaska Fairbanks, 2010. Pp. vi, 363. ISBN 9781555001124. $40.

Reviewed by MICHAEL DUNN, *Max Planck Institute for Psycholinguistics*

This collection of papers presents and discusses a landmark achievement in linguistics: Edward Vajda's first demonstration of a plausible genealogical link between languages of Eurasia and languages of the Americas (apart from the more recent circumpolar movements of Eskimo languages). Dramatic discoveries are rare in historical linguistics: as befits an old and often library-bound discipline, the pace of change in historical linguistics is slow, and most contributions are of the incremental type. In contrast, the Dene-Yeniseian hypothesis is a big step, and whatever the eventual consensus of its truth turns out to be, the proposal of Dene-Yeniseian is an important event in the development of historical linguistics and deserves all our attention.

The collection is based on papers presented at the Dene-Yeniseian Symposium held at the University of Alaska in February 2008. It starts with two useful and informative introductions: the editors' introduction giving the historical background to the Dene-Yeniseian hypothesis, followed by an introduction from the linguistic perspective by BERNARD COMRIE, which explains to a non-specialist readership the basic principles of the comparative method in historical linguistics and sketches the typological characteristics of the two component families, Yeniseian and Na-Dene. EDWARD J. VAJDA presents the case for the Dene-Yeniseian link in a sixty-six-page paper, which he follows with another paper in which he discusses Dene-Yeniseian in the context of other long-range comparative hypotheses, with generous acknowledgment of all possible intellectual precursors to the idea. Vajda makes clear his sympathy for the 'long ranger' tendencies in historical linguistics, but sticks to a mainstream approach in his analysis.[1] Part 2 of the collection contains interdisciplinary perspectives from human genetics (G. RICHARD SCOTT and DENNIS O'ROURKE), archaeology (BEN A. POTTER), areal phonology and Na-Dene historical linguistics (JEFF LEER), 'geolinguistics' (JAMES KARI), comparative kinship (JOHN W. IVES, SALLY RICE, and Edward J. Vajda), and folklore (YURI E. BEREZKIN, ALEXANDRA KIM-MALONEY). Part 3 of the collection is devoted to commentaries: a series of reviews of the evidence for the Dene-Yeniseian hypothesis by major figures in linguistics. While all of the authors of this section agree that the evidence for

[1] For instance, he cites the thirty-six Dene-Yeniseian etymologies presented in Ruhlen 1998, although it seems that only one has been validated by Vajda so far.