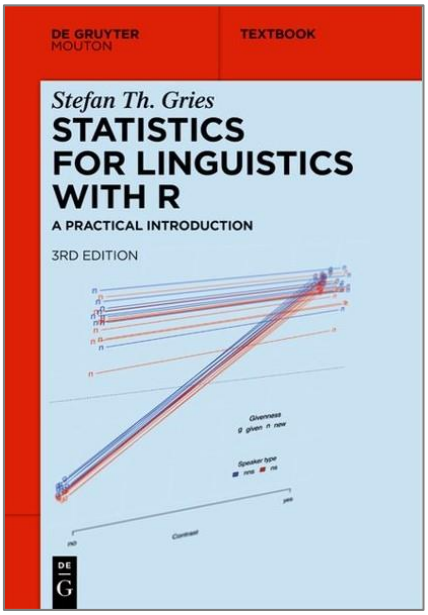
STUDIA UBB PHILOLOGIA, LXVII, 4, 2022, p. 519 - 522
(RECOMMENDED CITATION)

BOOKS

Stefan Th. Gries, *Statistics for Linguistics with R: A Practical Introduction*. Berlin/ Boston: Mouton de Gruyter, 2021, 495 p.



Having two previously published editions (2009 and 2013), the current edition (2021), *Statistics for Linguistics with R: A Practical Introduction*, revises and brings into discussion an updated part of the quantitative analysis methods adapted to linguistics and widely presented in their earlier versions. In order to shed light on the importance of statistics in linguistics, the author, Stefan Th. Gries, puts together some important background information that should be acquired by anyone interested in a quantitative type of research in the field of linguistics and languages. The author himself mentions that this particular book is different from other introductions in statistics because 'it has been written especially for linguists' (p. 1).

The textbook starts with an atypical, short, and informal *Introduction*, in which the author briefly justifies the need of a 3rd version of the book. He mentions the main reasons why he decided to work on a new edition, focusing, at the same time, on the major changes that he managed to bring in this latest version. Closely following the introduction, the *Table of Contents* presents the way in which the book is organised (chapters and subchapters). The total number of chapters is seven and they seem to have a logical sequence, starting with general fundamentals regarding statistics and going on with in-depth analyses and different approaches for specific linguistic data.

To start with, the first chapter is entitled *Some fundamentals of empirical research* and it focuses on creating a background for the reader when it comes to the quantitative type of research in linguistics, at the first glance, as the focus of the book is on linguistic data that can be statistically analysed. An important trait of this chapter is that the author discusses detailed information regarding the design that stays behind the models in quantitative studies. He presents clear methodology that should be followed in order to get relevant results or findings if

one approaches a quantitative method, focusing on some ways of organising the raw data before any type of analysis, for example, in a tabular format. Later on, the discussion continues with types of hypotheses in statistics and mathematics, correlating the methods used in these fields to the methods that can be addressed and adapted to linguistics as well. He talks about the importance of formulating the hypotheses accordingly in order to get a unified set of data and brings into discussion the difference between (i) non-directional or two-tailed hypotheses, (ii) directional or one-tailed hypotheses, and (iii) null hypotheses. Closely related, he mentions the use of variables as part of the method and differentiates between confounding and moderator variables. Putting together all these elements, they get us to the actual data collection and the crucial act of coding the information in a software in such a way we get the best comprehension of the methodology follow-ups. The end of the chapter focuses on how to address the data from a statistical point of view. It mentions the use of different techniques that evaluate the 'frequencies, distributions, averages/means, dispersions, or correlation coefficients' (p.29). In this part, the author also introduces some significance tests, emphasising *p* (probability) value which shows the significance limit. In the last few lines of the chapter, the author gives a short advice to his readers, talking about some inconsistencies that can occur when dealing with an empirical type of research.

The next chapter, *Fundamentals of R*, discusses the practical outcome of the software (*R Project for Statistical Computing*). It is a detailed description on how to install the software and all the particularities one should know before using it. This chapter is a more technical one and tries to give a clear representation of the terminology needed in order to have a better understanding of the matter. It clarifies the meaning of some essential terms, such as *functions*, *arguments*, *vectors*, *factors* or *data frames*, and the way in which we can generate, load, save, and work with these concepts. The purpose is obviously to enhance readers' skills and abilities when it comes to using the software accordingly to the principles of statistical research as correlated to linguistic research.

Going on, the third chapter focuses on *Descriptive statistics* and the specific analysis that could be performed in this type of approach. The dichotomy presented consists of univariate descriptive statistics and bivariate descriptive statistics. The former, univariate descriptive statistics, includes the categorical, ordinal, numeric variables as well as standard errors and confidence intervals examples. It also presents the distinction between the different central tendency aspect, focusing on the mode, the median, and the arithmetic mean. These concepts proved to be extremely important in linguistics as the researches often deal with a large amount of data that could be easily processed and analysed with the help of these tools (the mode can show the most frequent use of a specific structure; the median is meant to identify the middle of a sequence on a scale that questions the acceptability of a structure; the mean can easily represent the average scale when using a certain structure in language). The latter, bivariate descriptive statistics, focuses on the same concepts of categorical, ordinal, and numeric variables, however, this time represented as functions of different variables in the same model.

The following chapter is *Monofactorial tests* and sheds light on different distribution, frequency, dispersion, central tendencies, correlation, simple linear regression types of analyses that one can use to manage a statistical view on linguistics. In the introduction of the fourth chapter, the author briefly describes it as a chapter that shows 'how to decide which significance test to use' (p. 164), anaphorically presented in the first chapter of the textbook. In a step-by-step manner, a first question should target the kind of research one wants to conduct, for which the author presents three possible types: (i) descriptive; (ii) hypothesis-generating; (iii) hypothesis-testing. Based on each type of study conducted, a different approach would be needed. However, Gries decides to continue with the hypothesis-testing type of approach and elaborates a whole theory in order to give a comprehensive characterization of the types of tests suitable for this purpose. He also considers important to know the type of variables that one should involve in order to test the hypothesis as well as the number of variables included in the model. After the reader managed to gather all the information regarding the type of research, the author starts to discuss, in great detail, each type of analysis and its main advantages.

Perhaps the central chapter of this book, as stated by the author, chapter five is entitled *Fixed-effects regression modeling* and tackles concepts related to *multifactiorality*, focusing on both linear and binary logistic regression but also other models (multinomial regression and ordinal logistic regression). It is interesting to see a consistent amount of information that is organised in comprehensible graphic structures with actual examples related to languages and linguistics.

The current edition also presents two new chapters that were not tackled in the previous two editions. Firstly, it is chapter 6 which focuses on *Mixed-effects regression modeling*. It provides a case study for the linear mixed-effects regression, but it should be mentioned that the fixed-effects regression model is closely related to the mixed-effects one. For this, Grief exemplifies with a model design and the first step would be to build the fixed-effects regression structure (including fixed variables that do not change and are easily adapted to the nature of the study). The next step would be to focus on the mixed-effects regression structure, where one can include the random effects (variables that are context-dependent).

Lastly, it is the seventh chapter that presents *Tree-based approaches* and it introduces the regression trees. The author mentions this chapter as 'crossing from the domain of statistical modeling […] into that of machine learning' (p. 453), the latter domain being closely related to linguistics, especially in generative-transformational terms. The chapter mostly argues both in favor and against this model as this type of approach is still emerging and not yet a canonical type of analysis in linguistics. Even so, it is also brought into account an alternative emerging view and it refers to conditional inference tree. Different from the other tree-based approaches, it implies the use of *p*-value. In the end, the author concludes that the above-mentioned approach seems to be 'more intuitively interpretable' (472) as compared to other ways of visually organising the information, such as simple summaries or coefficients in tabular formats.

The last part of the book is represented by *References* and a note *About the author*. Surprisingly and atypically, the study does not have a general conclusion section, however, the author somehow concludes and summarizes the main ideas at the end of some chapters.

In the end, I would like to address a few observations regarding terminology, drawbacks but also favored traits regarding this textbook. First and foremost, the terminology used in the book is clearly explained even from the first two introductory chapters. However, when comparing this edition with its two previous editions (2009 and 2013), one gets to see a difference regarding the use of specific terms, for example, Chapter 4 was initially entitled *Analytical statistics*, however, in the current edition, we get a clearer understanding of the chapter from its title, namely, *Monofactorial tests*, which certainly represents a plus. Even so, I believe this motivation should be briefly addressed in the introduction of the textbook as well, alongside others, so that the reader figures out the inconsistency and correlates the two concepts accordingly (especially if the reader is already familiar with the previous editions). Secondly, it is worth mentioning that an improvement could be made with regard to abbreviations. Even if the main purpose is to explain different phenomenon and types of analyses throughout the book, I strongly believe an abbreviation list should be added at the beginning of the textbook just to make sure the readers have a unified comprehension of the technical words at least. Lastly, perhaps one of the traits of this book is the fact that it always provides sets of examples for each and every tackled issue, irrespective of the chapter or concept. More than that, something that I find extremely useful is that the author provides recommendations in terms of bibliography, advice regarding practical issues or even advice regarding the sets of tasks that one could complete in order to get familiar with the modelling designs that can be used later in actual research papers.

Overall, even if people that are not necessarily familiar with statistical approaches get to read the book, it could be enough to bring a good understanding regarding the correlation between the two fields as Gries uses a friendly 'interface'. With an increasing interest regarding the empirical and corpus-based research, the present book is surely becoming one important piece in the 'big puzzle picture' amongst the new generation of linguistic studies.

Nicoleta MANLUP
PhD student, Babeș-Bolyai University
Cluj-Napoca, Romania
Email: nicoleta.manlup@ubbcluj.ro