Introduction

1. Who is this book written for?

Statistics lies at the heart of scientific investigation. It helps the researcher to formulate and test theoretical hypotheses, making generalizations about a population of interest based on a limited sample. It is indispensable at all stages of the empirical cycle of research, from formulation of a hypothesis and data collection to data analysis and hypothesis falsification. Although statistics has played an important role in many hybrid linguistic disciplines, such as psycholinguistics, sociolinguistics, applied, computational and corpus linguistics, it is only recently that the awareness of its importance has reached the more traditional theoretical areas of linguistics. Slowly but surely, statistical methods are becoming a more common sight in linguistic teaching programmes, at conferences, in research articles and books.

This book is for you if you want…

- to learn how to operationalize and test your linguistic hypotheses
- to know which statistical method to choose in which situation
- to understand statistical terminology and participate in methodological discussion
- to get your message across with the help of clear and informative graphs
- to participate actively in making linguistics a more rigorous scientific discipline
- to get acquainted with a new programming language, R, and become a member of the dynamic international R community.

The book is intended mostly for researchers and students of usage-based or functional linguistics, although the methods are generic and can be applied in any field, regardless of one's theoretical persuasion. Without complicated jargon, but with detailed explanations, the book presents the most important statistical procedures and tests. The book also pays special attention to small-scale and non-normal data, which are so frequent in linguistic research. It shows how one can use non-parametric approaches and special techniques, such as bootstrap and permutation, to deal with small and irregular samples.

Most methods described in the book are well established in usage-based linguistics, such as logistic regression or distinctive collexeme analysis. However, some approaches are not yet mainstream. An example is Semantic Vector Spaces (Chapter 16), which originate in Computational Linguistics and can be used as a convenient tool to measure semantic relatedness of words, word forms or constructions.

Although some methods are more 'prototypical' than others in solving specific tasks, the suggested methods should not be regarded as the ultimate truth. Firstly, the development of tools that adequately solve theoretical problems in linguistics is still work in

progress, although some approaches have already become a *de facto* standard in particular fields (see examples in Newman 2011). Secondly, the methods are often complementary and allow one to see a phenomenon from different perspectives. The ultimate goal of the book is to encourage creativity by offering a set of classical and cutting-edge techniques that can be used to explore linguistic data.

Although R is a popular tool for extraction of information from corpora, this aspect is not covered in the textbook. For an introduction to corpus linguistics with R, see Gries (2009). We will not discuss the programming aspects of R, either. Those interested can consult textbooks, such as Chambers (2008) and Matloff (2011), as well as numerous online tutorials.

2. The quantitative turn in linguistics

In recent years, linguistics has been undergoing a quantitative turn, with statistical procedures gaining in popularity in many subfields, from usage-based morphology to Cognitive Semantics, and from phonology to discourse analysis. This has not always been the case, however. In fact, linguistics in the twentieth century has been dominated by idealist mentalist theories, which were alien to a rigorous empirical methodology. The most influential ones were (and still are, in some fields) Linguistic Structuralism and Generativism, which can be regarded as a modification of the former. Both assumed that the true object of linguistic investigation is some invariable structure, be it a system of oppositions, or innate linguistic competence. There was no need to resort to frequencies or probabilities since linguistic categories (at least, in one language), were assumed to be invariable, discrete and clear-cut. Every linguist could consider his or her linguistic knowledge a source of all necessary information about the entire language, which lead to the predominance of introspection and self-invented examples as the main type of evidence. As a consequence, statistics was thought to be unnecessary:

> Large groups of people make up all their utterances out of the same stock of lexical forms and grammatical constructions. A linguistic observer therefore can describe the speech-habits of a community without resorting to statistics.
>
> (Bloomfield 1935: 37)

> I think we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure. (Chomsky 1957: 17)

Today the situation is rapidly changing. On the one hand, these changes are in line with the recent theoretical shifts. One of them is connected with the spread of the usage-based approach, which originates in Langacker's work (1987). The main idea behind this framework is that linguistic knowledge is shaped by language usage. Importantly,

when speakers learn the language, they subconsciously analyse and store a vast amount of information about co-occurrence frequencies of words and constructions. When doing so, they act as 'intuitive statisticians' (Ellis 2006: 1). Frequency effects, which will be discussed in Chapters 5–7, play a crucial role in language use, acquisition and change. They are rooted in fundamental cognitive and social mechanisms. On the cognitive side, there is massive evidence that human categories have probabilistic structure and fuzzy boundaries, as shown by postclassical theories of categorization, such as Prototype and Exemplar Theories (cf. Chapter 19). From the social perspective, common linguistic categories (as well as shared conceptual structures) emerge as a result of linguistic alignment of speakers and hearers, which results in incremental strengthening of some representations and weakening of others over time. This process is the driving force of language evolution (Steels 2012). Obviously, the resulting inter- and intraspeaker variation can only be modelled statistically.

To implement the holistic approach to language use, variation and change, one needs multifactorial models. This approach reflects a general trend in linguistics labelled as 'recontextualization', which follows a long period of decontextualization in linguistic theory and practice (Geeraerts 2010). For example, one can test whether the speaker's choice between two or more linguistic alternatives in a particular context may be influenced by conceptual, social, stylistic, discursive, cultural and historical factors. Such models can be only created with the help of advanced multivariate methods, which are discussed in this book (see Chapters 12 to 14). Of course, when we deal with actual linguistic behaviour, there is always some degree of unpredictable variation, which is common in all types of social behaviour (Itkonen 1980: 350). This has to do with the complexity of the object of investigation, but should not discourage one from doing quantitative research.

On the other hand, even linguists who support theories that traditionally use introspection as the primary source of evidence begin to explore different types of data. As Kepser & Reis put it,

> [e]vidence involving different domains of data will shed different, but altogether more, light on the issues under investigation, be it that the various findings support each other, help with the correct interpretation, or by contradicting each other, lead to factors of influence so far overlooked. (Kepser & Reis 2005: 3)

Clearly, comparing different kinds of evidence from corpora, experiments, questionnaires, etc. requires an up-to-date methodological toolkit, which should necessarily include a broad range of statistical techniques.

Finally, the computer era has brought a huge number of freely available corpora, databases and other sources of linguistic evidence. Nearly every linguist nowadays has some experience with corpus data, even if this experience amounts to finding an unfamiliar expression in Google. Quantitative data analysis is indispensable when one needs to discover patterns in large data sets and corpora.

3. How to use this textbook

The textbook consists of four large parts. The first two chapters are preparatory. Chapter 1 introduces basic statistical concepts, whereas the main purpose of Chapter 2 is to help the reader get started with R. If you are new to R and statistics, this part of the book is indispensable.

The next two chapters are dedicated to the first descriptive analyses of quantitative (Chapter 3) and qualitative (Chapter 4) variables. They also discuss such basic notions as the mean, median, mode, standard deviation, proportions, and many more, and introduce a variety of standard plotting functions in R. This part of the book is also highly recommended for those who take their first steps in quantitative research.

The third and the largest part of the book explains main statistical tests and analytical statistics, from the *t*-test in Chapter 5 to the correlational analysis in Chapter 6 and linear regression in Chapter 7, followed by a discussion of different types of ANOVA in Chapter 8. Chapters 9–11 focus on association measures between two categorical variables and discuss collocational and collostructional methods, such as distinctive collexeme analysis (Chapter 11). Chapters 12 and 13 introduce logistic regression with binary and multinomial outcomes, respectively. These models are widely used in multifactorial probabilistic grammar and lexicology. Chapter 14 deals with additional classification tools, namely, conditional inference trees and random forests, which can be of help when regression analysis is not appropriate or the model is too complex to interpret. To get acquainted with some of the most popular hypothesis-testing techniques and measures of effect size, one may be well-advised to study Chapters 5, 6 and 9.

The last part of the book is dedicated to exploratory multivariate methods, which can be used to discover underlying structure in the data. Chapter 15 and 16 discuss some ideas behind distributional approaches to semantics: Behavioural Profiles and Semantic Vector Spaces, respectively. They also deal with various distance metrics and clustering techniques. Chapter 17 focuses on Multidimensional Scaling and shows how the method can be applied in variational linguistics. Chapter 18 is dedicated to Principal Components Analysis and Factor Analysis, which are used in multidimensional analysis of register variation. In Chapter 19, Simple and Multiple Correspondence Analysis are introduced and illustrated with case studies of lexical category structure and variation. Finally, Chapter 20 shows how one can visualize constructional change with the help of motion charts. Although the methods introduced in Chapters 15 to 19 are to some extent interrelated, each of these chapters can be studied on its own.

Every chapter contains R code and R output, including graphs. Executable R code is usually placed after the symbol '>' on a separate line, for example:

```
> a <-3
```

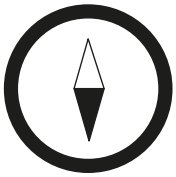
The '>' symbol should not be typed. Sometimes a short commentary is provided after the sign '#', for example:

```
> a <-3 # this is an example of a commentary
```

Appendix 1 summarizes the basic operations with a few most popular data types in R, whereas Appendix 2 offers an overview of numerous graphical parameters and plotting functions. It also introduces the package `ggplot2`, which enables one to produce high-quality graphs.

The book comes with a companion R package `Rling`, which is freely downloadable from the online platform for the textbook at http://dx.doi.org/10.1075/z.195.website (file *Rling_1.0.tar.gz*), see instructions in the file *read.me*. In addition, the online platform offers exercises and multiple choice questions that will help you master the basic statistical notions, specific methods and R code introduced in each chapter. The keys to the exercises and questions can be found online, as well as the R code for each chapter.

In every chapter you will see several boxes with additional information. The icons at the top of a box have the following meaning:

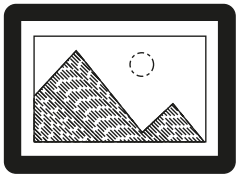


Additional information and reading suggestions.

A warning: it is easy to make a mistake.

Tips on writing up the results of a statistical test for publishing.

Practical recommendation or tip for fixing a problem.



Advice on how to create an enhanced version of a graph with the help of `ggplot2`.

The code provided in this textbook is based on R version 3.2.0. R is very dynamic and keeps changing and improving. This is why some coding details may change while this book is being published, as well as later on. Normally, when deprecated code is used, R gives a warning message and suggests an alternative. The readers would be well-advised to read such messages carefully and adjust the code accordingly.