Quantitative Methods in Linguistics

Keith Johnson

For Erin

Contents

Acknowledgments

This book began at Ohio State University and Mary Beckman is largely responsible for the fact that I wrote it. She established a course in "Quantitative Methods in Linguistics" which I also got to teach a few times. Her influence on my approach to quantitative methods can be found throughout this book and in my own research studies, and of course I am very grateful to her for all of the many ways that she has encouraged me and taught me over the years.

I am also very grateful to a number of colleagues from a variety of institutions who have given me feedback on this volume, including: Susanne Gahl, Chris Manning, Christine Mooshammer, Geoff Nicholls, Gerald Penn, Bonny Sands, and a UC San Diego student reading group led by Klinton Bicknell. Students at Ohio State also helped sharpen the text and exercises - particularly Kathleen Currie-Hall, Matt Makashay, Grant McGuire, and Steve Winters. I appreciate their feedback on earlier handouts and drafts of chapters. Grant has also taught me some R graphing strategies. I am very grateful to UC Berkeley students Molly Babel, Russell Lee-Goldman, and Reiko Kataoka for their feedback on several of the exercises and chapters. Shira Katseff deserves special mention for reading the entire manuscript during fall, 2006 offering copy-editing and substantive feedback. This was extremely valuable detailed attention - thanks! I am especially grateful to OSU students Amanda Boomershine, Hope Dawson, Robin Dodsworth, and David Durian who not only offered comments on chapters but also donated data sets from their own very interesting research projects. Additionally, I am very grateful to Joan Bresnan, Beth Hume, Barbara Luka, and Mark Pitt for sharing data sets for this book. The generosity and openness of all of these "data donors" is a high standard of research integrity. Of course, they are not responsible for any mistakes that I may have made with their data. I wish that I could have followed the recommendation of Johanna Nichols and Balthasar Bickel to add a chapter on typology. They were great, donating a data set and a number of observations and suggestions, but in the end I ran out of time. I hope that there will be a second edition of this book so I can include typology - and perhaps by then some other areas of linguistic research as well.

Finally, I would like to thank Nancy Dick-Atkinson for sharing her cabin in Maine with us in the summer of 2006, and Michael for the whiffle-ball breaks. What a nice place to work!

Design of the book

One thing that I learned in writing this book is that I had been wrongly assuming that we phoneticians are the main users of quantitative methods in linguistics. I discovered that some of the most sophisticated and interesting quantitative techniques for doing linguistics are being developed by sociolinguists, historical linguists and syntacticians. So, I have tried with this book to present a relatively representative and usable introduction to current quantitative research across many different subdisciplines within linguistics.[1]

The first chapter "Fundamentals of quantitative analysis" is an overview of, well, fundamental concepts that come up in the remainder of the book. Much of this will be review for students who have taken a general statistics course. The discussion of probability distributions in this chapter is key. Least-square statistics - the mean and standard deviation, are also introduced.

The remainder of the chapters introduce a variety of statistical methods in two thematic organizations. First, the chapters (after the second general chapter on "Patterns and tests") are organized by linguistic sub-discipline - Phonetics, Psycholinguistics, Sociolinguistics, Historical Linguistics, and Syntax.

This organization provides some familiar landmarks for students and a convenient backdrop for the other organization of the book which centers around an escalating degree of modeling complexity culminating in the analysis of syntactic data. To be sure, the chapters do explore some of the specialized methods that are used in particular disciplines - such as principal components analysis in phonetics and cladistics in historical linguistics - but I have also attempted to develop a coherent progression of model complexity in the book.

Thus, students who are especially interested in phonetics are well-advised to study the syntax chapter because the methods introduced there are more sophisticated and potentially more useful in phonetic research than the methods discussed in the phonetics chapter! Similarly, the syntactician will find the phonetics chapter to be a useful precursor to the methods introduced finally in the syntax chapter.

The usual statistics textbook introduction suggests what parts of the book can be skipped without a significant loss of comprehension. However, rather than suggest that you ignore parts of what I have written here (naturally, I think that it was all worth writing, and I hope it will be worth your reading) I refer you to the table below that shows the continuity that I see among the chapters.

[1] I hasten to add that, even though there is very much to be gained by studying techniques in natual language processing (NLP), this book is not a language engineering book. For a very authoritative introduction to NLP I would recommend Manning & Schütze's (1999) "Foundations of Statistical Natural Language Processing".

Hypothesis testing

Type of data		Predictor variables: Factorial (nominal)	Predictor variables: Continuous	Predictor variables: Mixed random and fixed factors
	Ratio (continuous)	T-test (ch 2 & 3) ANOVA (ch 4)	Linear regression (ch 2 & 3)	Repeated measures ANOVA (ch 4) linear mixed effects (ch 7)
	Nominal (counting)	C2 test (ch 5) Logistic regression (ch 5)	Logistic regression (ch 5)	Logistic linear mixed effects (ch 7)

Pattern discovery

Type of data		Type of pattern: Categories	Type of pattern: Continuous
	Many continuous dimensions	Principal components (ch 3)	Linear regression (ch 3) Principal components (ch 3)
	Distance matrix	clustering (ch 6) MD Scaling (ch 6)	
	Shared traits	Cladistics (ch 6)	

The book examines several different methods for testing research hypotheses. These focus on building statistical models and evaluating them against one or more sets of data. The models discussed in the book include the simple T-test which is introduced in chapter 2 and elaborated in chapter 3, analysis of variance (chapter 4), logistic regression (chapter 5), linear mixed effects models and logistic linear mixed effects models discussed in chapter 7. The progression here is from simple to complex. Several methods for discovering patterns in data are also discussed in the book (in chapters 2, 3, and 6) in progression from simpler to more complex. One theme of the book is that despite our different research questions and methodologies, the statistical methods that are employed in modeling linguistic data are quite coherent across subdisciplines and indeed

are the same methods that are used in scientific inquiry more generally. I think that one measure of the success of this book will be if the student can move from this introduction - oriented explicitly around linguistic data - to more general statistics reference books. If you are able to make this transition I think I will have succeeded in helping you connect your work to the larger context of general scientific inquiry.

A Note about Software

One thing that you should be concerned with in using a book that devotes space to learning how to use a particular software package is that some software programs change at a relatively rapid pace.

In this book, I chose to focus on a software package (called "R") that is developed under the GNU license agreement. This means that the software is maintained and developed by a user community and is distributed not for profit (students can get it on their home computers at no charge). It is serious software. Originally developed at AT&T Bell Labs, it is used extensively in medical research, engineering, and science. This is significant because GNU software (like Unix, Java, C, Perl, etc.) is more stable than commercially available software - revisions of the software come out because the user community needs changes, not because the company needs cash. There are also a number of electronic discussion lists and manuals covering various specific techniques using R. You'll find these resources at the R project web page (http://www.r-project.org).

Contents of the Book Web Site

The data sets and scripts that are used as examples in this book are available for free download at the publisher's web site - http://www.blackwellpublishing.com. The full listing of the available electronic resources is reproduced here so you will know what you can get from the publisher.

2. **Patterns and Tests**
Script: Figure 2.1
Script: The central limit function from a uniform distribution (central.limit.unif).
Script: The central limit function from a skewed distribution (central.limit).
Script: The central limit function from a normal distribution.
Script: Figure 2.5
Script: Figure 2.6 (shade.tails)
Data: Male and female F1 frequency data (F1_data.txt).
Script: Explore the chi-square distribution (chisq).

3. **Phonetics**
Data: Cherokee voice onset times (cherokeeVOT.txt).

Data: The tongue shape data (chaindata.txt).
Script: Commands to calculate and plot the first principal component of tongue shape.
Script: Explore the F distribution (shade.tails.df)
Data: Made-up regression example (regression.txt)

4. **Psycholinguistics**
Data: One observation of phonological priming per listener from Pitt & Shoaf's (2002)
Data: One observation per listener from two groups (overlap versus no overlap) from Pitt & Shoaf's study.
Data: Hypothetical data to illustrate repeated measures of analysis.
Data: The full Pitt & Shoaf data set.
Data: Reaction time data on perception of flap, /d/, and eth by Spanish-speaking and English-speaking listeners.
Data: Luka & Barsalou (2005) "by subjects" data.
Data: Luka & Barsalou (2005) "by items" data.
Data: Boomershine's dialect identification data for exercise 5.

5. **Sociolinguistics**
Data: Robin Dodsworth's preliminary data on /l/ vocalization in Worthington, Ohio.
Data: Data from David Durian's rapid anonymous survey on /str/ in Columbus, Ohio.
Data: Hope Dawson's Sanskrit data..

6. **Historical Linguistics**
Script: A script that draws Figure 6.1
Data: Dyen et al.'s (1984) distance matrix for 84 Indo-European languages based on the percentage of cognate words between languages.
Data: A subset of the Dyen et al. (1984) data coded as input to the Phylip program "pars".
Data: IE-lists.txt: A version of the Dyen et al. word lists that is readable in the scripts below.
Script: make_dist: This perl script tabulates all of the letters used in the Dyen et al. word lists."
Script: get_IE_distance: This perl script implements the "spelling distance" metric that was used to calculate distances between words in the Dyen et al. list.
Script: make_matrix: Another perl script. This one takes the output of get_IE_distance and writes it back out as a matrix that R can easily read.
Data: A distance matrix produced from the spellings of words in the Dyen et al. (1984) dataset.
Data: Distance matrix for eight Bantu languages from the Tanzanian Language Survey.
Data: A phonetic distance matrix of Bantu languages from Ladefoged, Glick & Criper (1971).
Data: The TLS Bantu data arranged as input for phylogenetic parsimony analysis using the Phylip program pars.

7. **Syntax**

Data: Results from a magnitude estimation study.
Data: Verb argument data from CoNLL-2005.
Script: Cross-validation of linear mixed effects models.
Data: Bresnan et al.'s dative alternation data.