Statistical Modeling in Linguistics

A systematic introduction with instructions for using R

Roland Schäfer

Dedicated to a randomly chosen individual.

Contents

Preface

What readers can expect from this book

This book has one primary purpose, which is to introduce students and practitioners to statistical methods in linguistics required to work with so-called *(Generalised) Linear (Mixed) Models* (GLMMs for short). This family of methods has been used productively in linguistics, both in corpus linguistics (for example in the modeling of alternations in Probabilistic Grammar) as well as experimental (psycho-)linguistics, where it tends to supersede traditional ANOVAs (for example in the analysis of reading time data). While GLMMs are advanced methods, this book is self-contained, leading readers directly from the introduction of basic notions about data and statistics to GLMMs. Other methods like so-called *descriptive statistics* and many statistical tests (like the t-test, χ^2-test, the simple ANOVA, etc.) are explained in passing.

The structure of each section within the chapters is the same, except in more general chapters like Chapters 1 and **??** (and more general sections in the other chapters). First, readers are introduced to a statistical problem. Second, the method used to solve this problem is explained in some detail, including calculations "by hand" where possible. At the end of each chapter, it is shown how to implement the methods discussed in the chapter in R. The section on R can be skipped if readers are not interested in doing statistics themselves, or if another software package is used. Wherever possible, general sections and R sections close with exercises, and it is highly recommended that readers do all of them.

The style of writing is (I hope) clear and concise, but I try to be comprehensive within the narrow focus of the book. While I try to avoid mathematical details where possible, this is simply not possible entirely. Practitioners who do not understand the *full* output of standard R packages for estimating GLMMs, for example, cannot make good and responsible use of them. Understanding this output, however, requires some understanding of the role of the variance-covariance matrices in so-called *mixed models* (see Chapter **??**). Hence, variance-covariance matrices are introduced.

After working through this book, readers should be able to use and understand

GLMMs (and the other methods introduced in earlier chapters) confidently. Even more importantly, they will be in a position to consult text books written by statisticians, which often seem aloof or even arcane to practitioners. The most useful text books to me are Fahrmeir et al. (2013), Fox (2016), Zuur et al. (2009), and above all Gelman & Hill (2006), all with different strengths and weaknesses. All of these references do not use linguistic examples, but there simply is no similar literature specifically written for linguists. Furthermore, they should know enough to consult online Q&A sites (see below) and blogs, which have become a valuable source of guidance for many users of statistical methods.

Finally, I want to point out that I often use simulated, artificial data for illustration purposes. Such made-up data sets make it easier to illustrate what kinds of data the method was designed for – and was *not* designed for. Also, many researchers do not make publicly available their data sets after publication, which limits a text book's author's options for choosing a data set for illustration.

"Am I qualified to write this book?"

This is the question that anybody who is not a trained statistician and sets out to write a book about statistics should answer after taking a deep, overly critical look in the mirror and preferably in a phase of very low self-esteem and almost pathological modesty. Other – usually more qualified – people will answer this question anyway after the publication, so seeking an honest answer before beginning to write the book is actually a form of self protection.

In a blog post entitled "Statistics textbooks written by non-statisticians: Generally a Bad Idea" (http://vasishth-statistics.blogspot.de/2016/11/statistics-textbooks-written-by-non.html), Shravan Vasishth of Potsdam University suggests that practitioners should not write this kind of text book. He quotes from a statistics text book written by a psychologist, which contains in one paragraph several misconceptions about statistical power and Type I and Type II errors (see Chapter 4). While I understand the frustration that motivated this blog post, I still think it is too extreme a measure to ban text books on statistics written by practitioners altogether. The typical errors and misconceptions in statistics text books, which rightly infuriate statisticians, are all related to the primordial sin of practitioners, which was blending the statistical paradigms of Ronald A. Fisher on the one hand and Jerzy Neyman and Egon Pearson on the other hand into the toxic slop referred to as Null Hypothesis Significance Testing (NHST; see Chapter 1). This has led practitioners to apply recipe-like statistical protocols without a proper understanding of what they are actually doing and – even worse – to make far-

reaching inferences which are not warranted the data. To make matters worse, this ritualised approach to "hypothesis testing" (NHST) is usually combined with Popperian falsificationist rhetoric (see also Chapter 1). Popperian falsificationism is declared to be the dominant methodological approach in *all* (Western) sciences, NHST is declared as a quantitative implementation of Popperianism, and thus anything we do with statistical tests is exemplary rational and objective science. Anyone who proposes such types of arguments demonstrates that he or she has probably never read even a text book about the Philosophy of Science (such as Chalmers 2013).

However, it is easy for individual practitioners to overcome this deficient state by reading a reasonably small selection of excellent and even accessible references written by – and this is crucial – philosophers and statisticians, not other practitioners. Statistics as handed-down to students by practitioners is often not second-hand but rather *n*-hand wisdom, where *n* is an arbitrarily high integer. Having done the necessary background reading, and having acquired a good understanding of the selection of methods that they use themselves, practitioners are, of course, qualified to write text books which provide basic introductions to selected methods and prepare students and colleagues for their own perusal of the original sources. Introductions by practitioners have the important function of introducing students and colleagues to a selection of methods relevant to their work. Furthermore, they make it easier to digest sometimes complicated matters by using examples from the field, and thus examples which readers can relate to.

Therefore, I decided to write this text book. To avoid getting things wrong, I adopted the following principles:

1. Try to be clear on your statistical philosophy and the kind of inferences warranted based on your methods! This is the hardest part for anyone who is neither a statistician nor a philosopher. In Chapter 1, I *try* to get things right, but I also encourage readers to be critical about that chapter, and to do some additional reading themselves.
2. Focus on *one* (family of) method(s)! In this case, I focused on Generalised Linear Models (GLMMs) and the methods needed to understand them. Readers will not find a catalogue of all possible statistical tests used in linguistics, nor short introductions to clustering methods, Principal Component Analysis, etc. The book starts from zero and leads directly to GLMMs, which do a lot of work both in corpus linguistics and experimental linguistics.
3. Make your explanations sufficiently detailed! All-in-one text books to statistics in any field (for example linguistics) on two hundred or less pages

will foster a culture of the recipe-like application of statistical methods which led to the current crisis of confidence in statistics in psychology, epidemiology, etc. (see Chapter 1).

4. Separate the introduction to the methods themselves from the introduction to their implementation in R! Again, this avoids training students to apply recipes in a software instead of understanding (to some degree at least) what they are doing. Also, this makes the book attractive for users of other software packages, which is a nice by-effect.
5. Make everything open access and available long before publication! This allows other practitioners and statisticians to audit the book before it is published, i. e., before severe harm can be done.

New types of sources and references

I take the open access and Creative Commons philosophy seriously. Also, I know that I have acquired a great deal of my knowledge of statistics not just by reading books and research papers, but from respected online sources. There are two main consequences for this book.

First, I reuse material from online sources which are released under a compatible Creative Commons license. Most prominently, I reused parts of the R Wiki Book at https://en.wikibooks.org/wiki/R_Programming instead of writing yet another introduction to R basics, data manipulation in R, etc. These parts will be marked appropriately, of course, and due credit will be given to the original sources. Second, I quote from blogs written by renowned statisticians, from the Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/), and from replies on the most important Q&A website about statistics, *CrossValidated* (https://stats.stackexchange.com/). Especially the blogs and contributions of the following people have been very inspiring and helpful, regardless of how often I reference them in this book (in alphabetical order):

- BenBolker (https://stats.stackexchange.com/users/2126/ben-bolker)
- Andrew Gelman (http://andrewgelman.com/)
- Debora G. Mayo (https://errorstatistics.com/)
- Richard Morey (http://bayesfactor.blogspot.de/)
- Stephen Senn (https://errorstatistics.com/tag/stephen-senn/)
- as well as many contributors on R-bloggers (https://www.r-bloggers.com/)

How this book was written

This book was created using XeLaTeX, which is in many ways superior to the old LaTeX. Mixing XeLaTeXand R code was made easy and elegant by knitr (https://yihui.name/knitr/), and I recommend anyone to use these three pieces of software to typeset their documents whenever they use statistics. Additionally, RStudio was used, also for illustration. Since I run an RStudio Server at https://webcorpora.org, I could work on this book in a perfectly platform- and machine-independent way. GitHub (https://github.com/rsling/smil) provided version control, and the available sources on GitHub can be used by anyone to see how it is done.

Acknowledgments

I am indebted to the authors of the R Wiki Book (https://en.wikibooks.org/wiki/R_Programming), from which I took (in adapted form) some of the very basic introductory material about R. In general, the statistics and R blogosphere and the *CrossValidated* community have contributed significantly to my understanding of statistics and the R programming language.

For valuable advice on several occasions, I would like to thank the statistical consulting team of Freie Universität Berlin, fu:stat (http://www.stat.fu-berlin.de/en/index.html).

I am also grateful to the participants of many university courses I taught at Freie Universität Berlin as well as workshops I taught elsewhere (including Göttingen University and Siegen University). Among the students in my statistics classes, Elizabeth Pankratz, Johanna Rehak, and Samuel Reichert provided the most valuable feedback through well-informed, inquisitive, and critical questions.

Finally, I would like to thank Felix Bildhauer for an ongoing discussion about statistical methods and their application in linguistics. This includes the "Dahlem statistics reading circle" and a shared repository of books and papers about statistics.

Abbreviations and symbols

Abbreviations

ANOVA	analysis of variance
CDF	cumulative distribution function
CLT	central limit theorem
cp.	ceteris paribus (all other things being equal)
iid.	independent and identically distributed
LM	linear model
LMM	linear mixed model
GLM	linear mixed model
GLMM	generalised linear mixed model
PDF	probability density function
VCOV	variance-covariance matrix

Symbols

Symbols are overloaded ad-hoc to denote either a (possibly indexed) value such as $s_x = 1$ (for "the population mean of variable x is 1") or a function such as $s(x) = 1$ where applicable.

	Mathematical symbols
$x \sim D$	x *follows* D (x a variable, D a distribution)
$\bar{x}$	sample arithmetic mean of x
$\tilde{x}$	sample median of x
$\hat{x}$	predicted value of x

	Letter-like symbols
α	alpha level
α_i	intercept i
β	beta level
β_i	first-level coefficient i

df	degrees of freedom
e	Euler constant
ϵ	observation-level error
f	frequency
F	F statistic (see ANOVA)
γ_i	second-level coefficient i
H	Kruskal-Wallis statistic
H_0	null hypothesis
H_A	alternative hypothesis
M_M	main hypothesis
IQR	inter-quartile range
$\mathscr{L}$	Likelihood
μ	population mean
μ_i	mean of modeled effect i
n	sample size
N	population size
O	Odds
p	proportion
P_i	the i-th percentile
Pr	probability
φ	dispersion parameter
Q_i	i-th quartile
r	sample covariance coefficient
r^2	coefficient of determination
R^2	multifactorial coefficient of determination
ρ	population covariance coefficient
s	sample standard deviation of x
s^2	sample variance of x
SE	standard error
SS	sum of squares
σ	population standard deviation
σ^2	population variance
U	Mann-Whitney statistic
χ^2	chi square statistic

Random distributions are denoted by bold-printed abbreviated names instead of the incoherent symbols sometimes used.

Bern	Bernoulli distribution
Exp	exponential distribution
F	F distribution
Norm	normal (Gaussian) distribution
t	t distribution
Unif	uniform distribution
Chisq	χ^2 distribution

1 Science, data, and statistics

1.1 Science, theories, and why we are not falsificationists

1.2 Data and inference

1.3 The importance of statistics

1.4 Why fake and simulated data – and how?

1.5 What is R, and why should you learn how to use it?

Testing:

Colorbox

Blabla

Iconcolorbox

Blabla

Sandwich

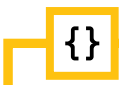
Blabla

Iconsandwich

Blabla

Framebox

Blabla



Iconframebox

Blabla

In this book, code listing are displayed as inline blocks such as the following simple code which simulates t-tests under the null pothesis in order to demonstrate that all p-values have equal probability under the null.

```r
# Set simulation parameters.
nsim  <- 1000
n     <- 100
meen  <- 0
stdev <- 1

# Data structure for results.
sims <- rep(NA, nsim)

# Simulations.
for (i in 1:nsim) {
  a <- rnorm(n, mean = meen, sd = stdev)
  b <- rnorm(n, mean = meen, sd = stdev)
  p <- t.test(a,b)$p.value
  sims[i] <- p
}
```

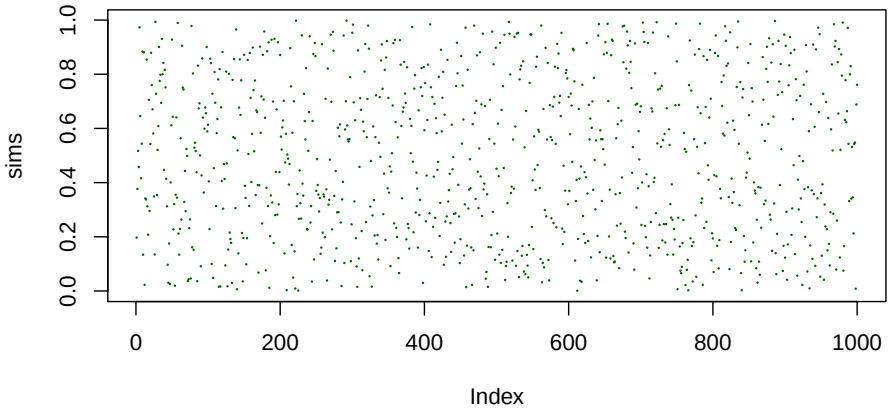


Figure 1.1: Scatterplot of p-values.

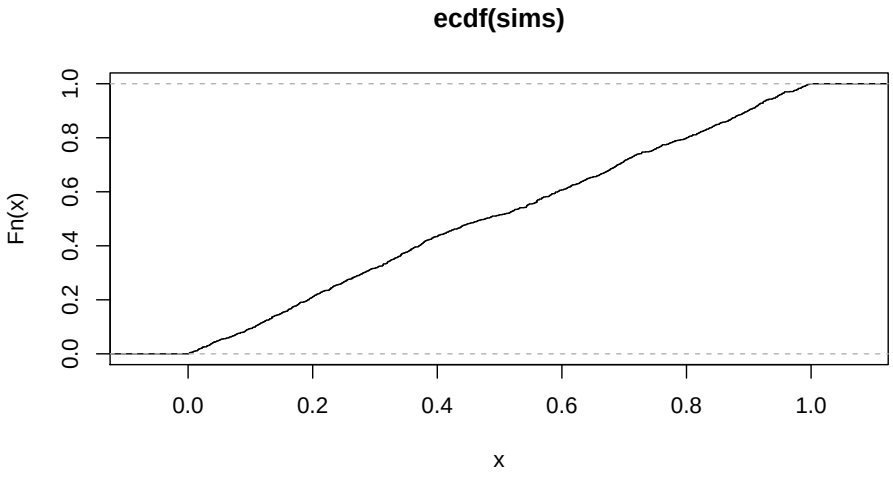


Figure 1.2: Empirical cumulative density distribution of p-values.

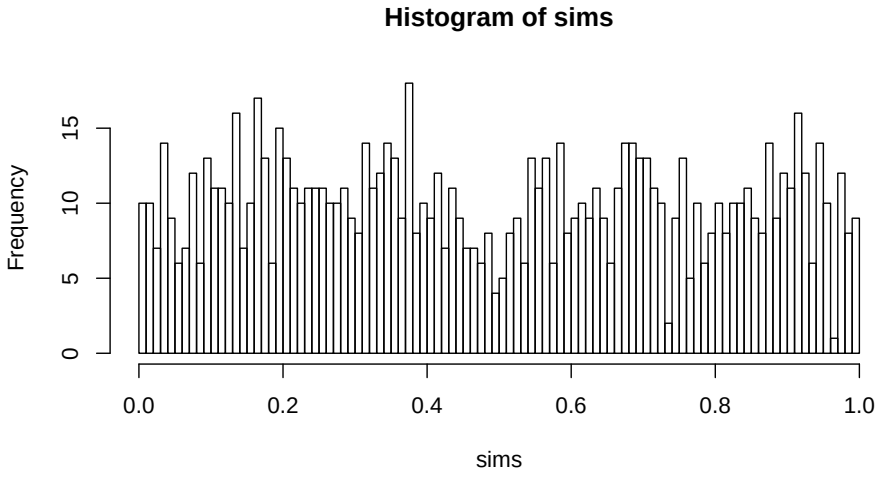


Figure 1.3: Histogram of p-values.

See Figure 1.2 for the cumulative density of p-values under the null in a series of 1000 t-tests. This was plotted using the following command:

```
plot(ecdf(sims))
```

2 Describing data

3 Visualising data

4 Tests

5 Models

6 Generalised models

7 Hierarchical models

8 Recent decelopments and outlook

8.1 Some recently proposed extensions and alternatives

8.1.1 Generalised Additive Models

8.1.2 Bayesian estimators

8.1.3 Neo-behaviourism

8.2 What is your statistical philosophy?

8.3 Open science and pre-registration

8.4 Further reading

8.5 Further R packages to look at

References

Chalmers, Alan. 2013. *What is this thing called science*. 4th edn. Maidenhead: McGraw Hill.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang & Brian Marx. 2013. *Regression – models, methods, and application*. Berlin etc.: Springer.

Fox, John. 2016. *Applied regression analysis & generalized linear models*. 3rd edn. London: Sage Publications.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer.