Shravan Vasishth · Michael Broe

The Foundations of Statistics: A Simulation-based Approach

Springer

The Foundations of Statistics: A Simulation-based Approach

Shravan Vasishth · Michael Broe

The Foundations of Statistics: A Simulation-based Approach

Springer

Shravan Vasishth
Department of Linguistics
University of Potsdam
Karl-Liebknecht-Str. 24-25
14476 Potsdam
Germany
vasishth@uni-potsdam.de

Michael Broe
Department of Evolution,
Ecology & Organismal Biology
Ohio State University
1304 Museum of Biological Diversity
Kinnear Road 1315
OH 43212 Columbus
USA
broe.1@osu.edu

ISBN 978-3-642-16312-8 e-ISBN 978-3-642-16313-5
DOI 10.1007/978-3-642-16313-5
Springer Heidelberg Dordrecht London New York

Cover design: WMXDesign GmbH

Cover image: Daniel A. Becker

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

SV dedicates this book to his son, Atri;
MB dedicates this book to his parents.

Foreword

Teaching the fundamental concepts and ideas of statistics is a relatively easy task when students have already completed courses in probability theory and calculus. Such students not only come well-prepared to an introductory statistics course, they are likely to follow additional, more advanced courses in mathematics and related areas. If so, many of the statistical techniques acquired remain relevant and are developed by further training.

For students without much of a background in probability or calculus, the situation is quite different. For many of us, just about everything is new, from integration to probability density functions, and from slopes of regression lines to random variables. Traditionally, introductory statistics courses targeting students from the arts and social sciences seek to explain the basic concepts and results of mathematical statistics, while omitting proofs. To survive such a course, one typically dutifully memorizes the equation for the density of the normal curve, the definition of the central limit theorem, and the sums of squares of a one-way analysis of variance. Once through this bottleneck, there is the safe haven of menu-driven software packages that will calculate everything one needs (and more). Sadly enough, many students will not come to appreciate the beauty of the fundamentals of statistics, and they will also remain somewhat insecure about what the basic concepts actually mean.

The approach taken by Shravan Vasishth and Michael Broe provides a much more interesting, exciting, and I believe lasting learning experience for students in the social sciences and the arts. (In fact, I believe that students in the sciences will also enjoy lots of the R code in this book). Simulation is a wonderful way of demystifying concepts and techniques that would otherwise remain abstract and distant. By playing around with simulated data, the reader can quickly get an intuitive feel for basic concepts such as the sampling distribution of the sample mean. With the tools provided in this book, the reader can indeed begin to explore the foundations of statistics, and will discover that statistics is actually fun and rewarding. Along the way, the reader will also acquire some basic programming skills, and will be well prepared to use the R software environment for statistical computing and

graphics. R is attractive not only because it is open source, and comes with thousands of add-on packages for specialized data analysis. R is also a great choice because of the many excellent books introducing statistical techniques in a reader-friendly and accessible way, with lots of example data sets. With this introduction, Shravan Vasishth and Michael Broe complement the existing R literature with a very useful and enjoyable hands-on introduction to statistics, doing analysis by synthesis.

Edmonton, October 2010 *R. Harald Baayen*

Preface

Statistics and hypothesis testing are routinely used in areas that are traditionally not mathematically demanding (an example is psycholinguistics). In such fields, when faced with experimental data in any form, many students and researchers tend to rely on commercial packages to carry out statistical data analysis, often without acquiring much understanding of the logic of statistics they rely on. There are two major problems with this approach. First, the results are often misinterpreted. Second, users are rarely able to flexibly apply techniques relevant to their own research – they use whatever they happened to have learnt from their advisors, and if a slightly new data analysis situation arises, they are unable to use a different method.

A simple solution to the first problem is to teach the foundational ideas of statistical hypothesis testing without using too much mathematics. In order to achieve this, statistics instructors routinely present simulations to students in order to help them intuitively understand things like the Central Limit Theorem. This approach appears to facilitate understanding, but this understanding is fleeting. A deeper and more permanent appreciation of the foundational ideas can be achieved if students re-run and modify the simulations themselves outside the class.

This book is an attempt to address the problem of superficial understanding. It provides a largely non-mathematical, simulation-based introduction to basic statistical concepts, and encourages the reader to try out the simulations themselves using the code provided on the course homepage http://www.purl.oclc.org/NET/vasishth/VB/. Since the exercises provided in the text almost always require the use of programming constructs previously introduced, the diligent student acquires basic programming ability as a side effect. This helps to build up the confidence necessary for carrying out more sophisticated analyses. The present book can be considered as the background material necessary for more advanced courses in statistics.

The vehicle for simulation is a freely available software package, R (see the CRAN website for further details). This book is written using Sweave

(pronounced S-weave), which was developed by Leisch, 2002. This means that LaTeX and R code are interwoven together.

The style of presentation used in this book is based on a course developed by Michael Broe in the Linguistics department of The Ohio State University. The first author (SV) was a student at the time and attended Michael's course in 2000; later, SV extended the book in the spirit of the original course (which was prepared using commercially available software). Both authors collaborated on the final text.

SV has used this book to teach linguistics undergraduate and graduate students at the University of Saarland, the University of Potsdam, and at the European Summer Schools for Language, Logic and Information held in Edinburgh (2005) and Bordeaux (2009). These courses have shown that the highly motivated student with little to no programming ability and/or mathematical/statistical training can understand everything presented here, and can move on to using R and statistics productively and sensibly.

The book is designed for self-instruction or to accompany a statistics course that involves the use of computers. Some of the examples are from linguistics, but this does not affect the content, which is of general relevance to any scientific discipline. The reader will benefit, as we did, by working through the present book while also consulting some of the books we relied on, in particular Rietveld & van Hout, 2005; Maxwell & Delaney, 2000; Baayen, 2008; Gelman & Hill, 2007.

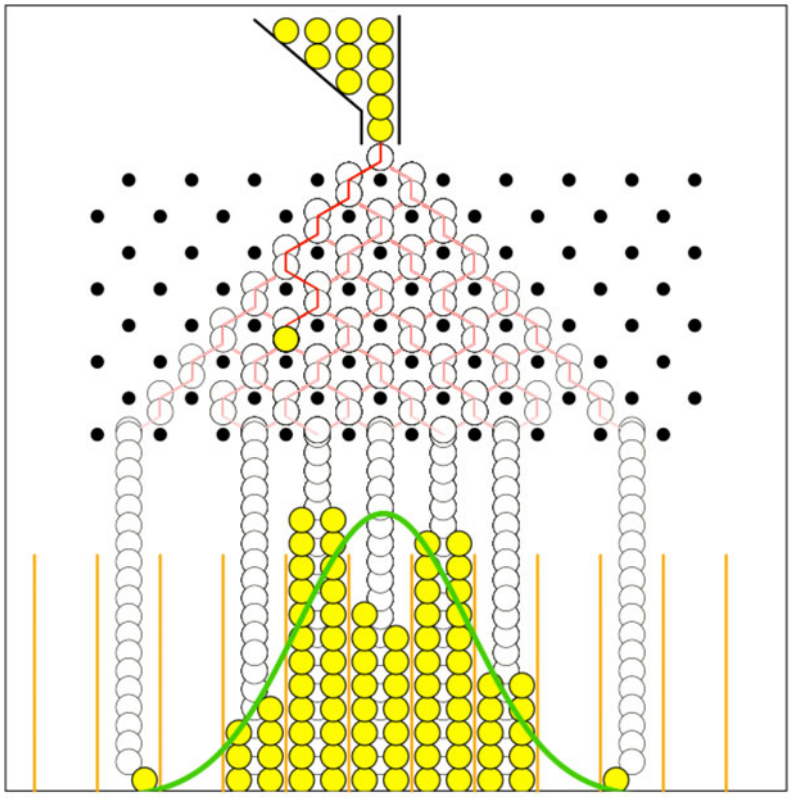
We do not aspire to teach R per se in this book; if this book is used for self-instruction, the reader is expected to either take the initiative themselves to acquire a basic understanding of R, and if this book is used in a taught course, the first few lectures should be devoted to a simple introduction to R.

After completing this book, the reader will be ready to use more advanced books like Gelman and Hill's *Data analysis using regression and multilevel/hierarchical models*), Baayen's *Analyzing Linguistic Data*, and the online lecture notes by Roger Levy.

A lot of people were directly or indirectly involved in the creation of this book. Thanks go to Olga Chiarcos and Federica Corradi dell Acqua at Springer for patiently putting up with delays in the preparation of this book. In particular, without Olga's initiative and efforts, this book would not have appeared in printed form. SV thanks Reinhold Kliegl, Professor of Psychology at the University of Potsdam, for generously sharing his insights into statistical theory and linear mixed models in particular, and for the opportunity to co-teach courses on statistical data analysis with him; his comments also significantly improved chapter 7. Harald Baayen carefully read the entire book and made important suggestions for improvement; our thanks to him for taking the time. Thanks also to the many students (among them: Pavel Logačev, Rukshin Shaher, and Titus von der Malsburg) who commented on earlier versions of this book. SV is grateful to Andrea Vasishth for support in every aspect of life.

MB thanks the students at The Ohio State University who participated in the development of the original course, and Mary Tait for continued faith and support.

We end with a note on the book cover; it shows a visualization of Galton's box (design by Daniel A. Becker). It is also known as the bean machine or quincunx. This was originally a mechanical device invented by Sir Francis Galton to demonstrate the normal distribution and the central limit theorem. The reader can play with the quincunx using the R version (written by Andrej Blejec) of the simulation, shown below (the code is available from Dr. Blejec's homepage and from the source code accompanying this book). The result of Dr. Blejec's code is shown below.



Berlin, Germany; and Columbus, OH *Shravan Vasishth*
October 2010 *Michael Broe*

Contents

Chapter 1
Getting Started

The main goal of this book is to help you understand the principles behind inferential statistics, and to use and customize statistical tests to your needs. The vehicle for this will be a programming language called R (google for 'CRAN,' which stands for the Comprehensive R Archive Network). Let's start with installing R and related software.

1.1 Installation: R, LaTeX, and Emacs

In order to use the code that comes with this book, you only need to install R. The latest version can be downloaded from the CRAN website. However, other freely available software provides a set of tools that work together with R to give a very pleasant computing environment. The least that you need to know about is LaTeX, Emacs, and Emacs Speaks Statistics. Other tools that will further enhance your working experience with LaTeX are AucTeX, RefTeX, preview-latex, and Python. None of these are required but are highly recommended for typesetting and other sub-tasks necessary for data analysis.

There are many advantages to using R with these tools. For example, R and LaTeX code can be intermixed in Emacs using noweb mode. R can output data tables etc. in LaTeX format, allowing you to efficiently integrate your scientific writing with the data analysis. This book was typeset using many of the above tools. For more on LaTeX, see the http://www.tug.org, the TeX Users Group website.

The installation of this working environment differs from one operating system to another. In Linux-like environments, most of these tools are already pre-installed. For Windows and Macintosh you will need to read the manual pages on the CRAN website.

After you have installed R on your machine, the second thing you need to do before proceeding any further with this book is to learn a little bit about R. The present book is not intended to be an introduction to R. For short,

S. Vasishth, M. Broe, *The Foundations of Statistics: A Simulation-based Approach*, DOI 10.1007/978-3-642-16313-5_1,

comprehensive and freely available introductions, look at the Manuals on the R homepage, and particularly under the link 'Contributed.' You should spend a few hours or even days studying some of the shorter articles in the Contributed section of the CRAN archive. In particular, you need to know basic things like starting up R, simple arithmetic operations, and quitting R. It is possible to skip this step and to learn R as you read this book, but in that case you have to be prepared to look up the online help available with R. For readers who want to start using the present book immediately, we provide a very basic introduction to R in Appendix B; the reader should work through this material before reading further.

1.2 How to read this book

We recommend that the book be read with an R session open on your computer and the code accompanying this book be kept open as an R file. The code used in this book is available from the homepage:

http://www.purl.oclc.org/NET/vasishth/VB/

You will get the most out of this book if you run the code as you read along, pausing to experiment with the code (changing parameters, asking yourself: "what would happen if I changed this setting?", etc.). Passive reading of the textbook will probably not yield much. Do not hesitate to re-read chapters! This material is best digested by revisiting it several times. The book chapters are intended to be read in order, since the later chapters build on concepts introduced in earlier parts of the book.

The accompanying website for the book contains (among other things):

1. A blog for asking questions that the authors or other readers can answer, or for submitting comments (corrections, suggestions, etc.) regarding the book that other readers can also benefit from.
2. Links to useful websites and other material relating to statistics.
3. Additional material that contains a more advanced treatment of some of the issues discussed in the textbook.

1.3 Some Simple Commands in R

We begin with a short session that aims to familiarize you with R and very basic interaction with data.

Let's assume for argument's sake that we have the grades of eleven students in a final examination for a statistics course. Both the instructor and the students are probably interested in finding out at least the maximum and

minimum scores. But hidden in these scores is much more information about the students' grades.

Assuming a maximum possible score of 100, let's first start up R and input the scores (which we just made up).

```
> scores <- c(99, 97, 72, 56, 88, 80, 74, 95, 66,
      57, 89)

 [1] 99 97 72 56 88 80 74 95 66 57 89
```

(When you type the above on the R command line you will not see R echo the contents of `scores`. You can type `scores` on the command line to print out the contents. Another way to print out the content of a command is to wrap it in parentheses: `(scores)`. We are able to print out the contents of `scores` above without using either of these methods because we are using Sweave with LaTeX, and Sweave allows the user to print when needed. See the Sweave homepage, http://www.stat.uni-muenchen.de/~leisch/Sweave/, for details)

Now we ask the following questions using R: (a) what's the maximum score? (b) what's the minimum?

```
> max(scores)

[1] 99

> min(scores)

[1] 56
```

We could stop here. But there is much more information in this simple dataset, and it tells us a great deal more about the students than the maximum and minimum grades.

The first thing we can ask is: what is the average or mean score? For any collection of numbers, their MEAN is the sum of the numbers divided by the length of the vector:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

The notation $\sum_{i=1}^{n}$ is simply an abbreviation for the statement that the numbers going from x_1 to x_n should be added up.

The mean tells you something interesting about that collection of students: if they had all scored high marks, say in the 90's, the mean would be high, and if not then it would be relatively low. The mean gives you one number that summarizes the data succinctly. We can ask R to compute the mean as follows:

```
> mean(scores)
```