

# 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 L<sup>A</sup>T<sub>E</sub>X 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.

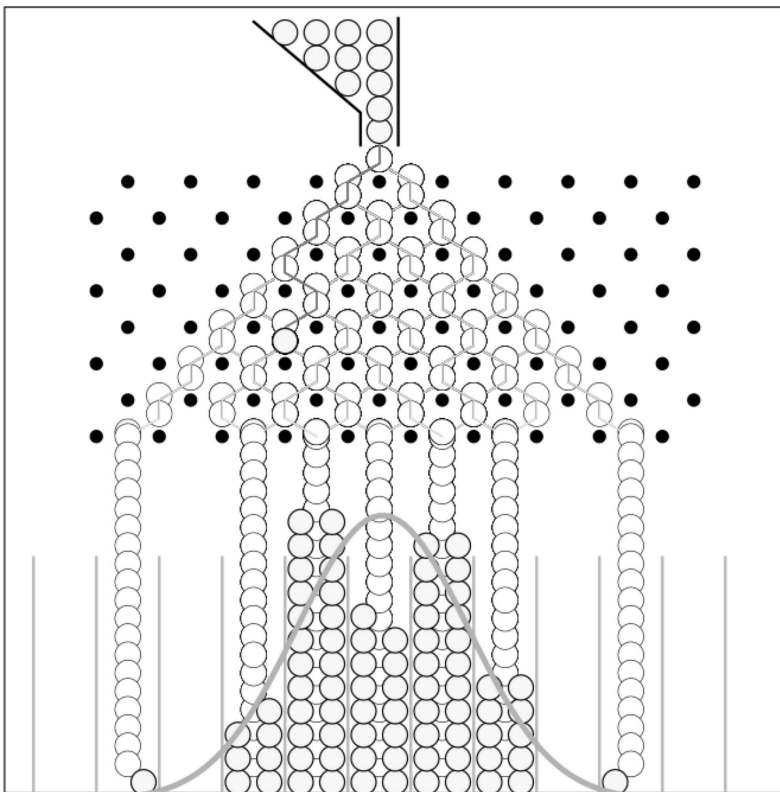
We do not aspire to teach R per se in this book; if this book is used for self-instruction, the reader is expected to either take the initiative themselves to acquire a basic understanding of R, and if this book is used in a taught course, the first few lectures should be devoted to a simple introduction 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  
October 2010

*Shravan Vasishth*  
*Michael Broe*

The Foundations of Statistics: A Simulation-based  
Approach

Vasishth, S.; Broe, M.

2011, XV, 178 p., Hardcover

ISBN: 978-3-642-16312-8