The Foundations of Statistics: A Simulation-based Approach, S. Vasishth and M. Broe, Springer-Verlag, Berlin Heidelberg, ISBN 978-3642163128 (hardcover, $50), xv+178 pages, by Christian P. Robert, Université Paris-Dauphine, Institut Universitaire de France, and CREST, Paris.

Table of contents

Readership: students with little background in probability or calculus, primarily in the social sciences and the art, ready to invest into R programming to explore probability and statistics. The book operates for both course work and self-instruction.

This book has been written by two linguists in order to teach statistics "in areas that are traditionally not mathematically demanding" at a deeper level than traditional textbooks towards building "the confidence necessary for carrying more sophisticated analyses" through R simulation. This is a praiseworthy goal, as simulation is indeed a perfect media to avoid "too much mathematics" and to build intuition. Instead, the book does not live up to expectations. Instead, there are statements there that show a deep misunderstanding of the subject.

Bypassing mathematical formulae sometimes leads to lengthy expository developments. For instance, the book spends four pages showing through an R experiment that "the sum of squared deviations from the mean are [sic] smaller than from any other number". It also entertains a permanent confusion between distributions and samples, between true parameters and their estimates. The book shies away from a proper definition of unbiasedness ("the mean of a sample is more likely to be close to the population mean than not" p.49), which leads to a poor justification of degrees of freedom and also to a most unfortunate section on "s is an Unbiased Estimator of σ", i.e. of the standard deviation. The t, chi-squared, and F densities are not introduced, while the normal density is defined as a "somewhat intimidating-looking function" (p.39) with a typo (since the exponential there is denoted by E), but the proper meaning of a density is never given. Similarly, the Central Limit theorem is not clearly stated and no mention is made of the Law of Large Numbers (although a connection is made in the summary,

p.63). In Section 2.3, the distinction between binomial and hypergeometric sampling is not mentioned, i.e. the binomial approximation is used with no warning. The introduction of the t distribution is motivated by the "fact that the sampling distribution of the sample mean can no longer be modeled by the normal distribution" (p.55). With such flaws in the material, it is difficult to recommend the book at any level, especially at the introductory one.

I am also dissatisfied with the way confidence and testing are handled in the book. The statement "We know that the value is within 6 of 20, 95% of the time" (p.49), which replicates the usual confidence fallacy, is found a few lines away from a warning about the inversion of confidence statements! A warning only detailed later stating that "its a statement about the probability that the hypothetical confidence intervals (that would be computed from the hypothetical repeated samples) will contain the population mean" (p.59). The book spends a large amount of pages on hypothesis testing, presumably because of its relevance within the field, however it is unclear a neophyte could gain enough expertise from those pages to conduct his own tests. Worse, statements like (p.75) $H_0 : \bar{x} = \mu_0$ show a deep misunderstanding of the nature of both testing and random variables. A similar confusion appears in the ANOVA chapter (e.g. (5.51) p.112).

The last chapters are about analysis of variance (5), linear models (6), and linear mixed models (7), all containing difficulties similar to the ones above. For instance, a series of hypotheses about the same data is processed as if they were independent (Section 5.1), true residuals are confused with estimated residuals in the regression chapter (p.138), a bootstrap on the response variable, rather than on the residuals, is used for testing a null hypothesis (p.139-140), and parameters are turned into random variables in the linear mixed model (p.152) without further explanation.

In conclusion, the book would have greatly benefited from a statistician's input or at least review before being published. As such, it cannot deliver the expected outcome to its readers and can neither train them towards more sophisticated statistical analyses. As a neophyte on linguistics, I cannot judge of the requirements of the field and of the complexity of the statistical models it involves. However, the most standard models and procedures should be treated with the appropriate statistical rigour, even at a minimal mathematical level. While the goals of the book were quite commendable, it seems to me it cannot endow its intended readers with the proper perspective on statistics.