The foundations of Statistics: a simulation-base approach, by Shravan Vasishth and Michael Broe

- **Hardcover:** 194 pages
- **Publisher:** Springer-Verlag, Berlin, New York
- **Language:** English
- **ISBN-10:** 3642163122

First, the title of this book is a misnomer, in that *The foundations of Statistics* is a light introduction to statistics for mathematically challenged students, using simulation, rather than any reflection on the foundations of our field. It is sadly plagued with errors that show the incomplete grasp of the authors have on their subject.

> *"We have seen that a perfect correlation is perfectly linear, so an imperfect correlation will be 'imperfectly linear'." page 128*

Those authors are Shravan Vasishth (Chair of Psycholinguistics and Neurolinguistics, Postdam, Germany) and Michael Broe (Department of Evolution, Ecology, and Organismal Biology, Ohio State University). Their purpose there is to teach statistics "in areas that are traditionally not mathematically demanding" at a deeper level than traditional textbooks "without using too much mathematics", towards building "the confidence necessary for carrying more sophisticated analyses" through R simulation. This is a praiseworthy goal, bound to produce a great book. However, and most sadly, I find the book does not live up to those expectations.

> *"Let us convince ourselves of the observation that the sum of the deviations from the mean always equals zero." page 5*

Besides the factual errors and foundational mistakes found therein, a puzzling feature of this book is the space dedicated to expository developments that aim at bypassing mathematical formulae, only to find this mathematical formula provide at the very end of the argument (as, e.g., the binomial pdf). Another difficulty is the permanent confusion between the sampling distribution and the empirical distribution, the true parameters and their estimates. If a reader has had some earlier exposition to statistics, the style and pace are likely to unsettle her. If not, she will be left with gaping holes in her statistical bases: for instance, the book contains no proper definition

of unbiasedness (hence a murky justification of the degrees of freedom whenever they appear), of the Central Limit theorem, of the t distribution, no mention being made of the Law of Large Numbers (although a connection is found in the summary, page 63). This is a strong gap, given the reliance on simulation methods throughout the book. The material therein thus does not seem deep enough to engage in reading Gelman and Hill (2006), as suggested at the end of the book. Having the normal density defined as the "somewhat intimidating-looking function" (page 39)

$$f(x) = \frac{1}{(\sigma\sqrt{2\pi})} E^{-((x-\mu)^2/2\sigma^2)}$$

and with a very unfortunate capital E certainly does not help. (Nor does the call to integrate rather than pnorm suggested to compute normal tail probabilities (pages 69-70), as it paradoxically requires more mathematical maturity. A minor point, admitedly.)

> *"The key idea for inferential statistics is as follows: If we know what a 'random' distribution looks like, we can tell random variation from non-random variation." page 9*

The above quote gives a rather obscure and confusing entry to statistical inference. Especially when it appears at the beginning of a chapter (Chapter 2) centred on the binomial distribution. As the authors seem reluctant to introduce the binomial probability function (pdf) from the start, they resort to an intuitive discourse based on (rather repetitive) graphs (with an additional potential confusion induced by the choice of an illustrative binomial probability equal to $p = 0.5$, since $p^k(1-p)^{n-k}$ is then constant in k). In Section 2.3, the distinction between binomial and hypergeometric sampling is not mentioned, i.e. the binomial approximation to the hypergeometric distribution is used without any warning. The fact that the mean of the binomial distribution $B(n, p)$ as np is not established and the one that the variance is $np(1-p)$ is not stated until the appendix, page 168. (Conversely, the book pends pages 36-39 showing through an R experiment that "the sum of squared deviations from the mean are smaller than from any other number".)

> *"The mean of a sample is more likely to be close to the population mean than not." page 49*

This quote is the concluding summary about the Central Limit theorem, following an histogram with 8 bins showing that "the distribution of the means is normal". It is itself followed by a section on "s is an Unbiased

Estimator of σ". This unfortunately fake result (here, s is the standard estimator of the standard deviation σ, which cannot be unbiasedly estimated) seems to indicate that the authors are unaware that the transform of an unbiased estimator is generally biased. The introduction of the t distribution is motivated by the "fact that the sampling distribution of the sample mean is no longer be modeled by the normal distribution" (page 55). With such fundamental flaws in the presentation, it is difficult to recommend the book at any level. Especially at the most introductory level where students or/and instructors have no other referential.

> *"We know that the value is within* 6 *of* 20*,* 95*% of the time." page 27*

I am also dissatisfied with the way confidence and testing are handled. The above quote, which replicates the usual fallacy about the interpretation of confidence intervals (since 6 is a realisation of a random variable), is found only a few lines away from a (correct) warning about the inversion of confidence statements. This warning is only detailed much later: "it's a statement about the probability that the hypothetical confidence intervals (that would be computed from the hypothetical repeated samples) will contain the population mean" (page 59). The book spends a large amount of pages on hypothesis testing, presumably because of the own interests of the authors, however it is unclear a neophyte could gain enough expertise from those pages to conduct her own tests. Worse, statements like (page 75)

$$H_0 : \bar{x} = \mu_0$$

show a deep misunderstanding of the nature of both testing and random variables, in the line of the earlier confusion between samples and distributions. A similar confusion appears in the ANOVA chapter (e.g. formula (5.51) on page 112). And as follows:

> *"The research goal is to find out if the treatment is effective or not; if it is not, the difference between the means should be 'essentially' equivalent." page 92*

The following chapters cover analysis of variance (5), linear models (6), and linear mixed models (7), all of which face fatal foundational deficiencies, similar to the ones pointed above. I quite understand that the authors wrote the book in a praiseworthy goal to reach to less sophisticated audiences and to the best of their abilities, however I remain amazed that the book did not undergo a statistician's review before being published. As is, it cannot deliver the expected outcome on its readers, i.e. cannot train them towards

more sophisticated statistical analyses. As a non-expert on linguistics, I cannot judge of the requirements of the field and of the complexity of the statistical models it involves. And I acknowledge that linguists, esp. students, do not have a strong mathematical background. Nor do I know of any available and valuable alternative at this level. However, I maintain that even the most standard models and procedures should be treated with the appropriate statistical rigor. In conclusion, it unfortunately seems to me the book cannot endow its intended readers with the proper perspective on statistics.

Further references

GELMAN, A. and HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.