Volume 22(2) May 2013

LIMNOLOGY AND OCEANOGRAPHY

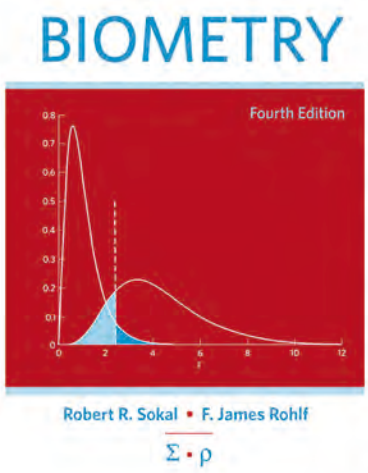
BULLETIN

ASLO

Association for the Sciences of Limnology and Oceanography

ROBERT R. SOKAL AND F. JAMES ROLF. 2012. **Biometry 4th Ed.** W.H. Freeman. ISBN-13 978-0-7167-86047, 937 pp, $115 (hardcover)

Reviewed by ***Stuart H. Hurlbert****, Department of Biology, San Diego State University, San Diego CA, U.S.A.; shurlbert@sunstroke.sdsu.edu*



Reader be forewarned: were it allowed the title of this review would be, "A readable but overblown, incomplete and error-ridden cookbook." Statistical analysis should be determined by the objectives, design and execution of a study. Unless these elements are appropriately incorporated into the subsequent statistical analysis, the arithmetic results can vary from meaningless to outright misleading. Understanding of that critical point is not reflected in this book (hereinafter referred to as SR). That is a key flaw that has been present since its first edition.

Shortly after publication of the first edition, biologist Leigh Van Valen (1970) reviewed it for *Science*. He had some quibbles but mostly praised it for having a discursive, easily readable style, excellent problems, and the best general treatment of correlation and regression he had seen. He concluded it is "easily the best introduction to biometrics, and perhaps to applied statistics generally, that is available." On the basis of that recommendation I and thousands of other biologists around the world went out and bought a copy. And biologists in the main have been singing the praises of successive editions ever since. The back cover of this new edition states it to be "the premier text in the field." Advertising by the company that distributes the software designed to accompany the book claims "this text has become the classic text on the application of statistics to biology since its initial publication in 1969."

A review by statistician Roger Mead (1982) of the second edition (not fundamentally different from the first, though 10 percent longer) expressed a radically different opinion. Mead concluded," It is compendious without providing clear exposition of the underlying principles. … I cannot recommend it as a textbook for a biologist wishing to understand statistical ideas, because of its cookbook style… it does have many merits even though, for me, these are clearly outweighed by its deficiencies." Among other deficiencies noted by Mead were the paucity of information on sampling design and experimental design and the fact that "in all sections of the book, the analysis of variance cart precedes the design horse." He also noted SR's mistaken assumption that "a proper explanation of the principles of statistical methods must be mathematical," its failure to acknowledge that many statisticians regard the whole idea

of correcting for "multiple comparisons" to be "faulty," and the large number of tests or procedures introduced briefly and superficially. Finally, Mead noted his concern "at the very low quality of much of the use of statistics by biologists" and observed that "the first edition of Biometry… is referenced much more frequently than any other by biologists whose statistics, in papers I referee, are incorrect."

Over the decades it has been my impression that the best applied statisticians have had, like Mead, as dyspeptic a view of SR as many biologists have had a worshipful attitude towards it. That the published record on the matter is not clear may be due primarily to the reticence of applied statisticians and to a scarcity of mathematical errors being sufficient to keep mathematical statisticians happy.

My review of this 4th edition of SR must be even more negative than was Mead's review of the 2d edition. Before getting into particulars, however, let me briefly tell readers what major changes in topical content have been made since the 3d edition. These include a new chapter on statistical power and sample size estimation, new sections on effect size, power, and sample size in several other chapters, and new sections on multiple comparisons, estimation of variance components and their confidence intervals, structural equation modeling, Akaike's information criterion and related criteria, and meta-analysis. This edition is now 937 pages long, up 21 percent from the 1st edition, and that despite earlier elimination of 95 pages of appendices on operation of desk calculators, Fortran statistical programs, and tabular guides to statistical methods. This 4th edition is truly super-compendious.

The book may be evaluated in light of the authors' claim that it presents "the minimum statistical knowledge required currently for a PhD in the biological sciences" (p. xv). The claim seems false on three grounds. First, any experienced, productive and numerate biological researcher can look at the smorgasbord of tests and procedures presented, often superficially, and note that the majority have been of no use to their own research and of little or no value to their ability to evaluate the literature critically. SR have been too anxious to include all sorts statistical gimmickery of very restricted, or even no, utility, perhaps impelled by a desire to seem up-to-date and thorough. Second, large areas of "statistical knowledge" understanding of which should be fundamental even for undergraduate students of statistics are left untreated by SR. Primary among these would be sampling and experimental design and the ongoing replacement of the hybrid Neyman-Pearsonian-paleoFisherian decision theoretic framework of significance assessment by a more logical neoFisherian one (e.g. Hurlbert and Lombardi 2009). Third, in a large number of places SR present not "statistical knowledge" but rather error, misinformation or bad advice.

The first of these grounds should be self-evident to any biological researcher and needs little documentation. All the statistical gimmickery occupies space that would have been better dedicated to the fundamental topics that have been ignored. The biologist can always run to a statistician for help in dealing with unusual data sets or problems. The expanded material

on methods for controlling set-wise type I error rates and on calculation of confidence intervals for standardized effect sizes are key examples of what I mean by gimmickery.

It is axiomatic that the design of a study should accord with a particular set of objectives and that analyses applied to resulting data must be appropriate to the design. For most of the examples they present, the design yielding the data is not explicitly described and sometimes cannot even be inferred. So one cannot be certain that the statistical procedures suggested are the appropriate ones.

For observational studies it is the basic concepts of sampling design that are relevant, and in these 937 pages there is not even a single small section on that topic. The preface (p. xiii) of SR seems to acknowledge the distinction between experimental and observational ("descriptive") studies, but later they often label particular examples of observational studies as "experiments," and incorrectly use terminology specific to experimental design to describe them.

Thus, on p. 344, SR describe as an "experiment" an observational study "to study the speed with which 10 rats repeatedly tested on the same maze, reach the end point," labeling the successive trials as the "treatment effect." On p. 406, SR describe as an "experiment" a study where dimensions of the lower face are measured on 15 girls when they are 5-years old and again, on the same girls, when they are 6. This is said to represent a "randomized-blocks design," with each girl being a "block" and age being the "treatment" factor. On p. 460, SR describe a putative "randomized blocks" design where temperature is measured at 10 depths in a lake on each of four successive days and "dates are considered blocks and the depths are the treatments." (The implication that limnologists would ever use any type of ANOVA in this way to assess change in a lakes' temperature profile is of course absurd). On p. 730, SR describe as an "experiment" the comparison of a hypothetical phenotype ratio to the observed phenotype ratios from eight replicates of a particular genetic cross of housefly types.

Of course the exercises above do not constitute experiments in the classical statistical sense of that term, and none involves blocks, treatment factors or treatment effects. Collectively these examples are more than sufficient to render the reader thoroughly confused about the distinction between experiments and observational studies, as well as the definitions of "block" and "treatment factor."

SR note that one benefit of the availability of high quality software for statistical analysis is that it now allows "researchers to concentrate on the more-important problems of proper experimental design and interpreting the statistical results…" (p. 33). But SR themselves do not assist this benefit. Aside from the numerous places where there is some discussion of randomized block designs (often with inappropriate examples, as noted above), the best this book can offer as a general introduction to experimental design is a two-page section ("Other designs") in the middle of the book (pp. 370-371). Two pages out of 937! There is no attempt to provide clear formal definitions for any of the key terms of experimental design -- such as experiment, experimental design, experimental unit, evaluation (or observational) unit, block, repeated measures, etc. – and there is no recognition of the tripartite structure of an experimental design. Incorrect definitions are sometimes imputed, e.g. "factorial" is implied to refer only to designs or analyses involving three or more treatment factors (p. 354).

Lack of clarity on the basic principles and concepts of experimental design not surprisingly leads SR to propose some unusually bad experimental designs. One involves a supposed randomized block design experiment on effects of a hormone treatment, where, in each of several cages, one control rat and one treated rat are placed (p. 353). The cage is considered the block. Such a set-up would violate the requirement that, for an unbiased estimate of a treatment effect, experimental units, even those in the same block, must be sufficiently isolated from or physically independent of each other that what transpires on one unit cannot influence what transpires on another. Other analyses of experiments described by SR constitute pseudoreplication (pp. 549, 758, 848) or would do so, if the implied analyses had been carried out (pp. 190, 371, 463, 796).

At the end of the two-page section on "Other designs," SR cite (p. 371) several works to which the reader might refer for further understanding of experimental design. Curiously most of the books cited are by biologists, most are as confused on the distinction between experimental and observational studies as are SR, and most contain plenty of other errors. Readers would be better advised to turn to clear and cogent experimental design books by professional statisticians, such as Cox (1958), Mead (1988), Hinkelmann and Kempthorne (2008), and Mead, Gilmour and Mead (2012).

With respect to the central topic of significance testing or assessment, SR adhere blindly to a rigid catechism that became fairly set in stone (though more for biologists than for statisticians) by the 1960s. Key doctrines in that catechism include: 1) an alpha must be specified prior to conduct of a significance test, and results of the test labeled as "significant" or "non-significant"; 2) if P is higher than alpha, then one should "accept" the null hypothesis, or at least assume the P value favors it over the alternative hypothesis; 3) if one has a notion or hope that the result will be in a particular direction, one should use a one-tailed test; and 4) when multiple comparisons or tests are conducted, the criterion for "significance" must be adjusted lest one risk too high a hypothetical set-wise (family-wise, experiment-wise) type I error rate.

Over the last half century many good statisticians, as well as many biologists and psychologists, have pointed out the illogical, counterproductive, even silly nature of those doctrines. SR seem completely unaware of even the existence of this large critical literature, let alone the cogency of the best of it. It is irresponsible to continue to push, without any caveats, this long outdated catechism onto our students and colleagues. Yet SR does this.

Of numerous other problems in SR, let me mention just a few:

- P. 83: A variance:mean ratio close to 1.0 does not indicate "cells are distributed approximately in Poisson fashion."
- P. 151: One- and two-tailed tests applied to the same data set "can lead to different conclusions" about what the data say, but only in the minds of the statistically naïve.

- P. 380: standardized effect sizes not only do not measure the "importance" of an effect, they usually do not even measure the magnitude of an effect in biologically meaningful terms.
- Pp. 413-422: The welter of detail on these tests for departures from normality or variance homogeneity obscures the unmentioned fact that it is not the "significance" of such departures (a matter strongly affected by sample sizes) that determines whether a given test will be robust in the face of such departures, but rather the magnitude of those departures. Graphical methods, rather than significance tests, are often recommended for assessing that, in conjunction with published studies on robustness of the specific procedure at hand.
- P. 427: SR state that when using log transformation and geometric means, reporting the back-transformed value of the standard error (s.e.) would be "misleading." This is not true so long as that back-transformed value is termed the "standard error factor" (s.e.f.) and reported with the geometric mean as gm x/÷ s.e.f. rather than as gm ± s.e.f. This will not apply, however, if, to deal with values of zero, a constant has been added to each datum before transformation.
- Pp. 430: When some zeros are in a data set and it is desired to use log transformation, the constant to be added to each datum should not necessarily be 1.0 but rather should be the lowest possible value of the variate possible, given the sampling and reporting protocols employed. If the distributions of the untransformed variate are at least approximately log normal, this will best favor normality and variance homogeneity in the transformed data sets. Negative characteristics in the logarithms pose no problem, so when there are positive variate values less than 1.0 there is no need to code the data "by multiplying by 10,000 or some higher power of 10…"
- P. 433: It is not true that count data (e.g. "insects on a leaf") "are likely to be Poisson … distributed" and that therefore square root transformation is likely to favor normality and variance homogeneity in data sets. Most organisms in most settings exhibit clumped or contagious distributions that are more highly right-skewed than are Poisson distributions. Log transformation will be a better choice for count data.
- Pp. 730-739: Ten pages are spent applying a "heterogeneity G-test" to phenotype ratios obtained from 8 replicate genetic crosses with the ultimate objective of comparing them collectively to a predicted phenotype ratio. The reader is left with very convoluted and poorly thought out advice about what to do if "heterogeneity" is detected. Replicate samples, like replicate experimental units, are never presumed to be identical and are always presumed to be "heterogeneous". Thus, the stated objective required only a one-sample t-test comparing the 8 sample ratios (perhaps converted to a percentages) to the predicted ratio (or percentage).
- P. 756: SR claim the phi coefficient of association for 2x2 contingency tables can range from -1 to +1. This is true, however, only for the special case when all marginal totals for the 2x2 table are the same, i.e. (a+b) = (c+d) = (a+c) = (b+d), in the conventional notation. For any given set of marginal totals, the only association coefficient that can range from -1 to +1 is C8, a modified version of a coefficient proposed by Lamont Cole (Hurlbert 1969).
- P. 793: It is not true that the experimental units to be used in a study must be "selected randomly" from all those "available to the investigator."

This edition of SR, like earlier editions, does generally provide clear, readable, step-by-step instructions for carrying out the tests and procedures it recommends. Doubtless most of those instructions are correct. But also like earlier editions, this one provides few conceptual frameworks for understanding when and why, if at all, a particular procedure is needed or appropriate, it gives negligible attention to the basic concepts of sampling and experimental design, it does not recognize that design and statistical analysis are two distinct aspects of a study, it muddies the distinction between experiments and observational studies, it dedicates large amounts of space to matters of trivial importance, and it contains many more simple errors than are listed here. There truly is nothing to recommend SR as a course text. Experienced researchers, statisticians, and editors with a critical eye might benefit from having a copy at hand as a compendium of mostly accurate 'recipes'. It will also help them understand one origin of the discombobulated statistical approaches found in many manuscripts that are crossing and will continue to cross their desks. After three further decades of heavy reliance on SR by biologists since Mead's review of the 2d edition, this book alone must account for a large fraction of the uncritical statistical thought and poor statistical advice found both in the primary literature of biology as well as in other biologist-authored statistics books.

Perhaps this review has judged SR by too high a standard. There are some better, less encyclopedic statistics textbooks available. But none is sufficiently neoFisherian, sufficiently free of errors, sufficiently clear, and sufficiently focused on principles to be considered ideal for a modern introductory statistics course. We can hope there are a few good writer-statisticians warming up in the wings.

REFERENCES

Cox, D.R. 1958. *Planning of Experiments.* Wiley, New York.

Hinkelman, K. and O. Kempthorne. 2008. *Design and Analysis of Experiments, vol. I: Introduction to Experimental Design, 2d ed.* Wiley-Interscience, New York.

Hurlbert, S.H. 1969. A coefficient of interspecific association. *Ecology* 50:1-9.

Hurlbert, S.H. and C.M. Lombardi 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici* 46:311-349.

Mead, R. 1982. Review of *Biometry*, 2d ed., by R.J. Sokal and F.J. Rohlf. *Biometrics* 38:863-864.

Mead, R. 1988. *The Design of Experiments.* Cambridge University Press, Cambridge.

Mead, R., S.G. Gilmour, and A. Mead 2012. *Statistical Principles for the Design of Experiments.* Cambridge University Press, Cambridge.

Van Valen, L. 1970. Statistics in biology Review of *Biometry*, 1st ed., by R.J. Sokal and F.J. Rohlf, *Science* 167:165.