Book Reviews

Experiments in Ecology
by A.J. Underwood

Cambridge University Press, 1996.
£22.95 (US$34.95) pbk; £65.00
(US$85.00) hbk (xviii + 504 pages)
ISBN 0 521 55696 1 pbk; 0 521 55329 6
hbk

This tome is ambitious in its scope and unique in its approach. It is about logic, statistics and design. Its examples are ecological but all are hypothetical. This is not an analysis of ecological concepts or case studies. The ecology is secondary.

Despite its title, the book is more focused on analysis of variance than on manipulative experimentation. Underwood uses 'experiment' in its oldest, pre-statistics sense, to denote any empirical observation or study designed to answer a question or test an hypothesis. Thus he dedicates much space to observational studies and sampling design, mixing this material into the same chapters that discuss manipulative experiments and experimental design *sensu stricto*. The distinction between manipulative experiments and mensurative experiments, or hypothesis-testing observational studies, Underwood considers a 'distraction' (p. 16), a dismissal that will set many statisticians' teeth on edge.

The book additionally covers *in extenso* the basic concepts of parametric statistics (100+ pages) and Underwood's particular vision of the 'logical structure' appropriate to the testing of research hypotheses. Underwood has strong and controversial opinions on many statistical topics. With this book he courageously puts them into one convenient package for our scrutiny. Here, I organize my comments under the three themes – logic, statistics and design – that Underwood lays out for the book, and add a few remarks on terminology also.

Underwood advocates a logical framework that he claims is 'well-used and of long-standing' and 'in widespread use in ecology' (p. 4). We should hope this is not true. The framework combines the falsification procedure of Karl Popper and the decision-theoretic framework of Jerzy Neyman and Egon Pearson in a way unlikely to have been acceptable to any of these fellows. And the framework is completely antithetical to R.A. Fisher's concept of significance testing.

In the typical situation, when the test of a null hypothesis (H_0) yields a low P value, we all agree that one has reasonable grounds for rejecting H_0 and has 'something to talk about'. On the other hand if the test yields a high P value, we have little to talk about. In particular, the high P value is not evidence in favor of either H_0 or the alternative hypothesis (H_a). A high *P* value is a recommendation only for indecision with respect to the truth of H_0. This is elementary. But in Underwood's 'logical framework', a high *P* value indicates that H_0 should be 'retained' and that H_a is 'clearly wrong', 'disproven', and 'falsified' (p. 17).

The fundamental difficulty seems to be Underwood's belief that it is possible and desirable to conflate two distinct logical frameworks into one – a general framework for testing of research hypotheses and the narrow framework of significance testing by which particular data sets bearing on a research hypothesis are evaluated, one by one.

Underwood applies his belief that null hypotheses can be confirmed not only to significance testing of substantive questions, e.g. treatment effects, but also to diagnoses that are used to determine the type of analyses used in testing the substantive questions. For example, if tests for heterogeneity of variances, for presence of interaction, or for differences among experimental units yield high *P* values, it is reasonable, in Underwood's view, to conclude that variances are homogeneous, interaction is zero, and experimental units are identical (or that we are very close to these conditions) – and to modify procedures for testing the substantive questions accordingly. These recipes are flawed. Their consequences for type I error are not spelled out and can be great. For example, when there are multiple samples per experimental unit and real differences among experimental units go 'undetected' (as low power would often cause them to be), Underwood's recipe leads directly to pseudoreplication: in tests for treatment effects, *P* values will be biased, usually downward, to an unknown degree. In other areas Underwood also advocates many procedures that are as widely accepted as they are without logical foundation. These include the fixing of alpha, the probability of a type I error, the use of 1-tailed tests when the direction of a result is predicted, and the fixing of set-wise type I error rates when multiple comparisons are made. But these are large battles that cannot be fought here.

Especially disconcerting is his treatment of the notion of independence. He dedicates many pages to discussing different types of non-independence and their consequences for statistical tests. Among scientists, there is much confusion over the concept of statistical independence and various intuitive notions about physical independence. The former pertains only to the little epsilons in our models and can be evaluated only in reference to both a data set *and a specified hypothesis*. If we take a set of random samples of bug density from each of two plots, the 'errors' (epsilons) will possess the statistical independence needed for testing the H_0: *no difference between plots.* But, in the case where one plot has been sprayed with an herbicide and the other kept as a control, these errors will *not* possess the statistical independence required for testing the H_0: *no difference between treatments.*

When we adopt an imprecise shorthand that uses 'independent' as a qualifier for 'data', 'samples', 'designs', 'events', and so on, it is easy to fall into a trap, and Underwood does. Here are three examples. He states (p. 169) that in order to test, for plants in a field, the H_0: *species A and B are equally abundant*, mean density must be estimated in a separate set of quadrats for each species. In an experimental example (p. 177), he objects to the pairing in space of experimental and control plots, and disparagingly labels a standard randomized complete block design as a 'non-independent design'. In both these examples he tries to demonstrate via simulations that the claimed 'non-independence' leads to biased *P* values. The 'bias' detected, however, is an artifact of his applying the wrong test (one-way ANOVA), one that ignores the pairing or blocking in the 'non-independent' designs. In the third example (p. 319) the objective is to compare male and female mortality rates under different, experimentally established population densities. Underwood asserts that the need for independence requires the establishing of two sets of experimental plots at each density so that, even though both sexes are present in each plot, male and female mortality can be measured in separate plots. This is completely unnecessary.

With respect to experimental design, Underwood also makes some controversial claims. I list a few, though there is not space to evaluate them here. *Caveat emptor.* They include: 'there are rarely situations where it is possible to make valid comparisons between a single experimental and a single control group of replicates' (p. 132); unequal replication of treatments should generally be avoided (pp. 156, 380); randomized block designs usually should be used only if they have within-block replication of treatments (p. 391); most split-plot designs have some special susceptibility to 'spatial non-independence among treatments or replicates because the different units are so close' (p. 400).

A consistent, precise terminology for their field has eluded statisticians so far. This is not a particular concern for Underwood, so for many terms he simply passes on the pre-existing confusion. His usage of 'experiment' and 'independence' have been mentioned. He tends to use 'sample', 'sample unit', and 'experimental unit' interchangeably. His definitions of 'effect size' and 'factorial experiment' are unconventional. And, unforgiveably, he uses four terms I have proposed (pseudoreplication, pseudofactorialism, mensurative experiment, demonic intrusion) but misdefines each one. Mensurative experiment, for example, he equates with a 'goodness-of-fit' test (p. 21)! And pseudoreplication he terms a 'neologism' for the broad, vague concept of 'confounding' (p. 245), which is equivalent to calling 'kookaburra' a neologism for 'bird'.

It is his bad luck that a hardcore Popper–Neyman–Pearsonian like Underwood has had to have his book reviewed by a mellow Fisherian like myself. Our disagreements concern more than the logical framework for hypothesis-testing, of course. Underwood's sharply defined positions should stimulate much salutary discussion among ecologists and statisticians.

Stuart H. Hurlbert