

PSEUDOREPLICATION IN HERMIT CRAB SHELL SELECTION EXPERIMENTS: COMMENT TO WILBER

Emili Garcia-Berthou and Stuart H. Hurlbert

Statistical methods are a tool intensively used in the biological sciences. However, their use is often careless, as several reviews have shown (Underwood, 1981; Hurlbert, 1984; Hurlbert and White, 1993). One of the most common errors is pseudoreplication. This was committed in 27% of field ecological experiments analyzed by Hurlbert (1984), in 41% of experiments on freshwater invertebrate zooplanktivores analyzed by Hurlbert and White (1993), and in 12% of field ecological experiments published in 1991–1992 and analyzed by Heffner et al. (1996).

In a short note, Wilber (1993) attempted to show that pseudoreplication committed in previous studies on hermit crab shell selection did not affect their conclusions. We believe there are problems in the statistical analyses and interpretations in this note. Our main criticisms are summarized below. We should first clarify the aim of our note. This is not so much a critique of Wilber's research but of current statistical practice in general. The quality of Wilber's statistical analyses is similar to that of much of the biological literature. Previous reviews sometimes have each surveyed a large number of published works. There are two reasons we choose to analyze in detail a single note: (1) it directly addressed pseudoreplication and wrongly concluded that although present it was not a problem, and (2) it exemplifies several statistical pitfalls common in the biological literature. Obviously, the quality of an investigation depends on more than good statistical practice and most of the papers with misuses still contain useful information. Our goal in this note is to emphasize the need to avoid pseudoreplication and to provide a concrete instructional example.

Pseudoreplication was defined by Hurlbert (1984) as "the use of inferential statistics to test for treatment effect with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent" or, more exactly, as the statistical treatment of multiple evaluation units in an experimental unit as if they each represented a separate experimental unit (Hurlbert and White, 1993). Pseudoreplication would be present in previous studies of crab shell selection if the variation among crabs within an experimental unit (generally a tank or aquarium containing multiple crabs) was used as an estimate of variation among experimental units within a treatment. Though some of those studies (e.g., Mitchell, 1976; Wilber, 1990) involved preference trials, which generally do not fall under the heading of manipulative experiments and thus do not involve experimental units, the concept of pseudoreplication can be extended to these and other types of observational studies (Hurlbert and White, 1993).

Our first criticism is that Wilber (1993) commits sacrificial pseudoreplication (Hurlbert, 1984; Hurlbert and White, 1993) in the second paragraph of the Results section, where his main statistical analysis was reported. The objective of his analysis of covariance (ANCOVA) for effect of hermit crab length on length of gastropod shell selected was to compare the slopes and intercepts of regression functions of crabs individually-tested with those for crabs tested in groups. The experiment consisted of four sets of 20 crabs

each tested individually and two sets of 20 crabs tested in groups. Even if there were no significant differences (in regression lines) between the two sets of group-tested crabs, the differences among crabs within those two replicates cannot be used as within-treatment variation. The pooling carried out prevents the variation among the sets, or experimental units, from being used as it should be in the significance test of the difference between individually- and group-tested crabs. The demonstration that experimental units are not significantly different is not evidence that variation among experimental units is zero and is not a legitimate justification for pooling.

An appropriate alternative analysis would be to directly compare the mean slopes and mean intercepts of the regression functions for the two treatments (type of crab) by means of a t-test. A regression function would be calculated separately for each of the two sets of group-tested crabs and for each of the four sets of 20 individually-tested crabs. The two slope values thus obtained for group-tested crabs would be compared with the four slope values obtained for the individually-tested crabs by a t-test; and the same would be done for the intercept values.

The use of simple F tests (ratio of two variances) to compare error mean squares (MS) of four regression analyses is also problematic. The use by Wilber (1990, 1993) of error MS instead of correlation coefficients is insightful. However, to compare the variances among more than two groups, some other test (e.g., Levene, Cochran, or Bartlett) that takes into account the number of groups being compared is required. Moreover, the P values of these F tests were wrong. They should be doubled (i.e., $F_{19, 19} = 1.67, P > 0.2$; $F_{19, 19} = 1.25, P > 0.5$; $F_{39, 79} = 1.05, P > 0.5$) because a two-tailed test is called for (e.g., Sokal and Rohlf, 1995: 190; Zar, 1999: 137). We would also recommend reporting the degrees of freedom with statistic values, e.g., to clarify the ANCOVA model applied.

Apart from the other statistical problems, the meaning of high P values was also misinterpreted. In the first paragraph of the discussion, it was claimed to have shown that pseudoreplication did not affect the conclusions of previous experiments. As a part of the argument, it was considered that a "non-significant" P value proves or substantiates the null hypothesis. This is a common but gross statistical error. Interpretation of P values is very asymmetrical. A "significant" result (low P value) is strong evidence against the null hypothesis, i.e., we can consider that the null hypothesis is false. In contrast, a nonsignificant P value provides no grounds for rejecting the null hypothesis but also no grounds for accepting or retaining it. When sample sizes are small or variances large, it is common to get high P values even when the null hypothesis is false. Power analysis has been suggested in these circumstances, although the fixing of arbitrary values for α and effect size pose further difficulties.

Finally, Wilber (1993) claimed to have shown that "inferences drawn from previous experiments are not limited by the potential for non-independence among replicates." Even if his analyses of his own data set and his conclusions regarding it were correct, much care should be taken before extrapolating such conclusions to other studies carried out in different contexts and with different material. If as the author suggests, testing crabs in groups limits the applicability of results because of aggressive interactions, they should be tested individually. This is an aspect of the experimental design beyond statistics. But in any case, pseudoreplication is a different problem, a statistical error that is easy to avoid.

ACKNOWLEDGMENTS

P. Petraitis and three anonymous reviewers improved the manuscript with constructive criticisms. A stay of EGB at San Diego State University was partially funded by the University of Girona.

LITERATURE CITED

- Heffner, R. A., M. J. Butler IV and C. K. Reilly. 1996. Pseudoreplication revisited. *Ecology* 77: 2558–2562.
- Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54: 187–211.
- _____ and M. D. White. 1993. Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. *Bull. Mar. Sci.* 53: 128–153.
- Mitchell, K. A. 1976. Shell selection in the hermit crab *Pagurus bernhardus*. *Mar. Biol.* 35: 335–343.
- Sokal, R. R. and F. J. Rohlf. 1995. *Biometry*. 3rd ed. Freeman, New York.
- Underwood, A. J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. *Oceanogr. Mar. Biol. Ann. Rev.* 19: 513–605.
- Wilber Jr., T. P. 1990. Influence of size, species and damage on shell selection by the hermit crab *Pagurus longicarpus*. *Mar. Biol.* 104: 31–39.
- Wilber, P. 1993. Pseudoreplication in hermit crab shell selection experiments: does it occur? *Bull. Mar. Sci.* 52: 838–841.
- Zar, J. H. 1999. *Biostatistical analysis*. 4th ed. Prentice Hall, Upper Saddle River, New Jersey.

DATE SUBMITTED: September 15, 1998.

DATE ACCEPTED: June 23, 1999.

ADDRESS: *Department of Biology, San Diego State University, San Diego, California 92182; PRESENT ADDRESS: (E.G.B.) Dept. Ciències Ambientals, Universitat de Girona, E-17071 Girona, Catalonia, Spain.*