Ecology, 88(1), 2007, pp. 56–62

SIMPLICITY AND COMPLEXITY IN ECOLOGICAL DATA ANALYSIS

PAUL A. MURTAUGH[1]

Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

Abstract. I argue that ecological data analyses are often needlessly complicated, and I present two examples of published analyses for which simpler alternatives are available. Unnecessary complexity is often introduced when analysts focus on subunits of the key experimental or observational units in a study, or use a very general framework to present an analysis that is a simple special case. Simpler analyses are easier to explain and understand; they clarify what the key units in a study are; they reduce the chances for computational mistakes; and they are more likely to lead to the same conclusions when applied by different analysts to the same data.

Key words: analysis of variance; correlated data; mixed-effects model; P value; repeated measures; statistics.

INTRODUCTION

I can easily test the hypotheses by simple t tests, but want something more "elegant" that will fit well with a "better" journal.

—Oregon State University professor requesting statistical help

This comment reflects an attitude that is all too common in ecology, and probably many other fields: the more complicated and technical-sounding the data analysis, the more compelling will be the conclusions from that analysis. In my view, simple statistical arguments are more persuasive than complicated ones, since one can more easily follow the train of logic and verify that there are no weaknesses obscured by elaborate statistical manipulations. This is apparently a minority view, as evidenced by an abundance of needlessly complicated and confusing statistical methodology in modern ecological literature.

The statistical analyst tries to quantify in an objective way the extent to which a set of data supports or refutes particular hypotheses about the population(s) from which the data came. Statistical theory has been developed to reduce the data to their essence, e.g., by identifying sufficient statistics, which are more compact than the original set of data, yet contain all of the information pertinent to the hypotheses of interest. Inferences based on these statistics are usually as direct and powerful as possible.

It is often not necessary to develop explicit models for observations made below the level of the experimental or sampling unit in a study (so-called "subsamples"), since unit-specific summaries will be the foundation for statistical inference. Avoiding detailed descriptions of subsamples simplifies the presentation of the analysis and clarifies what the effective sample size really is. Simple models are more likely to convey the key features of the data to other scientists, and they increase the chances that different analysts will reach the same conclusions.

It is natural, as a scientific field matures, for mostly descriptive work to be supplanted by more rigorous observational and experimental protocols, with a corresponding increase in the use of statistical inference. But some ecologists seem to go overboard in the detail and complexity of their statistical presentations. For example, Schluter (1994) presents 14 *P* values in describing a study based on two divided ponds, and Stewart-Oaten (1996) and Murtaugh (2000) use 15 and 14 equations, respectively, in their discussions of models of time series from two ecological units. Following are two other examples of analyses that I think could have been presented more simply and clearly.

All of the modeling and calculations in this paper were done with R (R Development Core Team 2005).

EXAMPLE 1: SIZES OF ZOOPLANKTON IN EXPERIMENTAL PONDS

Murtaugh (1989) studied the size and species composition of zooplankton in six experimental ponds at Cornell University, three of which had reproducing populations of planktivorous fishes and three of which were fishless. I measured the body lengths of thousands of zooplankters, and tested for a difference in body size between the fish-containing and fishless ponds. In the following sections, I present two ways of comparing the sizes of macrozooplankton between pond types: the first approach was used in the original paper, and the second I came up with more recently. I illustrate the approaches using a set of 288 zooplankton sizes selected randomly from the published data, shown in Fig. 1.

Manuscript received 22 February 2006; revised 2 August 2006; accepted 14 August 2006. Corresponding Editor: A. M. Ellison.

[1] E-mail: murtaugh@science.oregonstate.edu

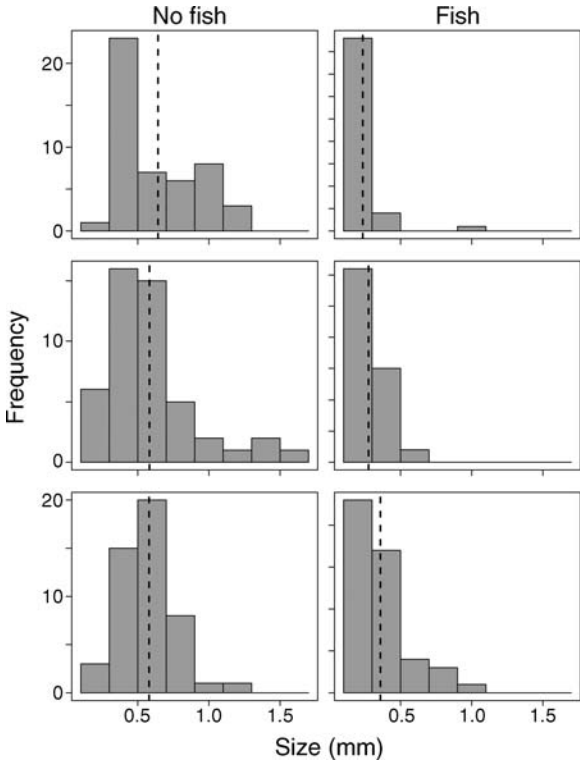

FIG. 1. Sizes of macrozooplankton in six experimental ponds at Cornell University, Ithaca, New York, USA (Murtaugh 1989): three ponds without fish (left column), three with fish (right), 48 zooplankton measurements per pond. The vertical dashed lines show the average sizes in each pond, used in the second analysis of the data.

Analysis 1: Nested analysis of variance

Here is an excerpt from Murtaugh (1989) explaining this method:

"I ran a nested analysis of variance (ANOVA) on the lengths of zooplankton in the six ponds—the effect of pond on body size is nested within the effect of predation treatment. Note that, since the study ponds can be thought of as a sample from a larger population of ponds that might have been used, 'pond' is a random effect nested within the fixed effect of fish ... Because of the complications involved in nested ANOVA's with unequal sample sizes (Sokal and Rohlf 1981), I fixed the number of replicates per pond at the value for the pond with the smallest number of length measurements available ...; appropriately-sized samples were obtained from the larger sets of measurements by random sampling without replacement."

Table 1 summarizes the nested analysis of variance on the current subset of the zooplankton size data. We have strong evidence that mean sizes differ between the fish treatments ($F_{1,4} = 55.78$, $P = 0.002$). It is worth noting that this analysis could also be called one-way analysis of variance with subsampling (see Kuehl 2002:159).

Analysis 2: Two-sample t test

Comparing the mean sizes in the fish-containing ponds (0.642, 0.582, and 0.580 mm) to the mean sizes in the fishless ponds (0.235, 0.275, and 0.358 mm), a two-sample t test yields $P=0.002$ ($t_4=7.469$). We have strong evidence that mean sizes differ between the fish treatments.

Comparison of approaches

Obviously, analysis 2 is simpler and easier to understand. Furthermore, it is mathematically equivalent to analysis 1: the F statistic in analysis 1 is the square of the t statistic from analysis 2 (see Appendix A). Nevertheless, the first analysis somehow seems more impressive than the second. If someone had told me in 1989 that my nested ANOVA could be restated as a two-sample t test, I would have been dismayed by the mismatch between the effort involved in data collection and the simplicity of the subsequent analysis. And I'm guessing that some readers would have been skeptical that an analysis of thousands of size measurements in six ponds could be summarized so simply.

In spite of its veneer of respectability, I see no advantages to analysis 1. It is harder to understand; it takes longer to explain; and it gives details about a component of variation (zooplankters within pond) that is not directly relevant to the hypothesis test of interest. If there is interest in the nature of the distribution of zooplankton size within ponds, graphical approaches (like the histograms in Fig. 1) are much more informative than an extra line in an ANOVA table.

This is not to say that the effort of measuring thousands of zooplankters was in vain; the larger the number of measurements contributing to each pond-specific mean, the more precise is our estimation of the mean. That is, as the number of measurements per pond increases, the variability of the pond means within each treatment will decrease, making it easier to detect a possible fish effect on zooplankton size. In this case, interestingly, retrospective power calculations (not shown) suggest that I would have been nearly as likely to detect the fish effect if I had made only a tiny fraction of the number of measurements that I actually did. A small amount of pilot data, along with some guidance about optimal allocation of sampling effort in the face of multiple components of variation (e.g., see Kuehl 2002: 163), could have saved me considerable time and effort.

Some might contend that the simplicity of the presentation of analysis 2 obscures assumptions that are implicit in the analysis. In fact, the assumptions behind the two analyses are identical, viz., that the six pond means are independent and normally distributed random

TABLE 1. Nested ANOVA table for the zooplankton size data.

Source	df	SS	MS	F	P
Fish	1	7.009	7.009	55.782	0.0017
Pond within fish	4	0.503	0.126	2.709	0.0305
Residual	282	13.080	0.046		

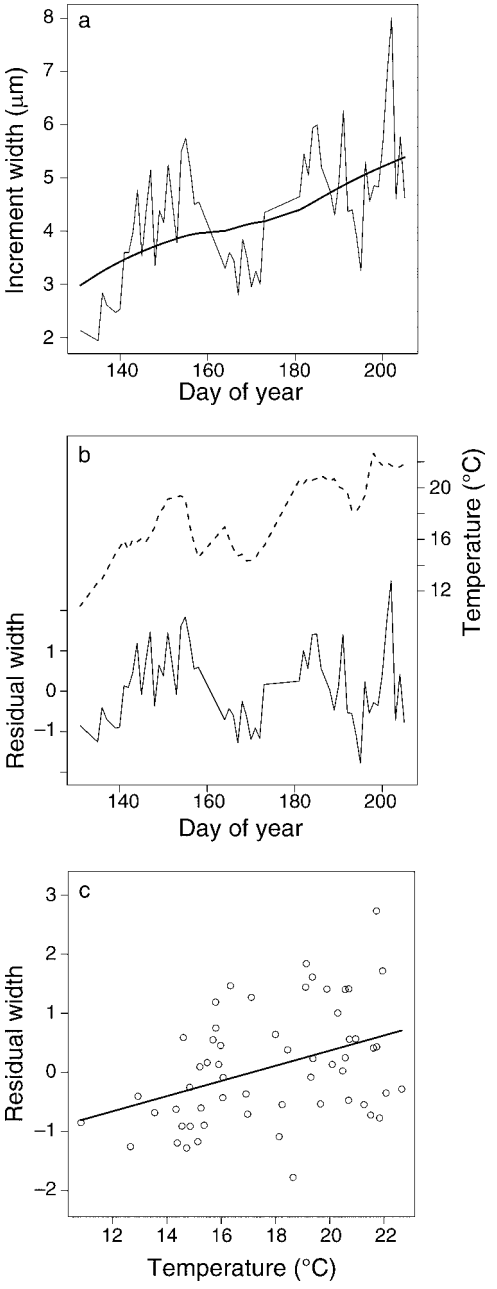

FIG. 2. Illustration of the processing of data from a single Lost River sucker (Terwilliger et al. 2003), showing (a) width of daily otolith increments (in μm) vs. the date on which they were deposited, with a "lowess" fit superimposed; (b) increment-width residuals from the above fit vs. date, and lake temperature (in °C) vs. date; and (c) increment-width residuals vs. lake temperature, with a least-squares regression line superimposed.

variables with equal variances. The choice of whether to summarize the results as a t test or as an F test from an ANOVA table is a matter of stylistic preference.

The difference between the two approaches comes in when extra attention is paid to the within-pond component of variation: this could be merely the use of a fancy name (nested ANOVA, or ANOVA with subsampling), or the analyst might proceed with another F test, comparing between-pond to within-pond variation in zooplankton size. The latter test requires an additional assumption—that the individual size measurements are normally distributed—and it is not directly germane to the test for an effect of fish on body size. The t test does not require this extra assumption, because the Central Limit Theorem ensures that means of 48 observations will be approximately normally distributed, regardless of the distribution of the individual observations.

EXAMPLE 2: ASSOCIATION OF TEMPERATURE WITH FISH GROWTH

These data are from a study of the associations of environmental variables with the daily growth rates of suckers in Upper Klamath Lake, Oregon (Terwilliger et al. 2003). Daily growth increments of the otoliths from juvenile fish were related to a variety of variables measured in the lake. Here I focus on a set of 53 Lost River suckers collected in 1995, and consider the association of temperature with the width of otolith increments. The total number of measured increments was 4996.

Fig. 2 illustrates the steps used in processing the data from this study: time series of increment widths for individual fish were "de-trended" by fitting lowess curves (Fig. 2a; Cleveland 1981), and residuals from those fits were then compared to ambient temperatures on the days corresponding to the growth increments (Fig. 2b). Fig. 2c shows one way of visualizing the relationship between residual increment width and temperature for an individual fish. The following two analyses are ways of summarizing these relationships over the sample of 53 fish, in an effort to make broader inferences about the association between temperature and increment growth rate in the larger population of fish.

Analysis 1: Linear mixed-effects (LME) model

This is the approach used by Terwilliger et al. (2003), following my recommendation. A general model for the residual increment width of fish i (in μm) at age t_j (in days) can be written as

$$y_{ij} = \eta_i + \beta_i x_j + \varepsilon_{ij} \quad (1)$$

where η_i is a random intercept associated with fish i; x_j is the temperature (in °C) on the date that this fish was age t_j; β_i is the change in increment width for fish i associated with a one-degree increase in temperature; and ε_{i1}, ε_{i2}, …, ε_{in_i} are sequential errors for fish i.

To fit a specific variant of the model in (1), a number of decisions must be made (e.g., see Pinheiro and Bates 2000, Venables and Ripley 2002):

1) Should we allow each fish to have its own relationship between increment width and temperature (i.e., model β_i as a random effect), or should we assume that all fish share the same relationship (i.e., let $\beta_1 = \beta_2 = \cdots = \beta_{53} = \beta$, a fixed effect)?

2) What is the nature of the serial correlation of the errors within fish (ε_{i1}, ε_{i2}, …, ε_{in_i})? Possible models

include an autoregressive process of order p, AR(p); a moving-average process of order q, MA(q); a combination of autoregressive and moving-average processes; and a continuous-time autoregressive process, CAR(1), in which the time position variable can be any continuous variable. Once a model is chosen, it must be "tuned" to the data—for example, one must select values of p or q, or, in the case of the CAR(1) model, one must decide how the time position variable is to be specified (in the fish example, it could be Julian days corresponding to the growth increments, or simply the order of observations within each fish).

3) After the model and the within-group correlation structure have been specified, one must decide whether to estimate parameters using maximum likelihood (ML) or restricted maximum likelihood (REML). The two approaches, which differ in the way that variance components are estimated, can give quite different results when the sample sizes and number of groups are small (Venables and Ripley 2002).

4) What do we do if the estimation algorithm fails to converge for our data set? Most users of linear mixed-effects models have faced this problem at one time or another.

Analysis 2: Individual regression lines

In this alternative approach, the increment-width residuals are regressed against ambient temperature separately for each of the 53 fish, resulting in 53 least-squares estimates of the association between increment width and temperature, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{53}$ (Fig. 3). Since it is unlikely that one fish's growth increments "influence" those of another fish, it is reasonable to assume that the 53 regression coefficients are statistically independent. Furthermore, the least-squares regression estimates are unbiased for the underlying population mean, no matter what the correlation structure of the errors within fish (e.g., see Diggle et al. 2002: 59).

The variances of the fish-specific $\hat{\beta}$'s will vary, depending on the number of observations made for each fish and the temperatures recorded on those dates. Unbiased estimates of those variances are available from the output for the 53 fish-specific regression fits. The optimal summary is therefore a weighted average of the fish-specific regression coefficients, with weights proportional to the reciprocals of the squared standard errors from the individual fits:

$$\hat{\bar{\beta}} \equiv \sum_{i=1}^{53} w_i \hat{\beta}_i$$

$$w_i = \frac{1\Big/\left[\mathrm{SE}\left(\hat{\beta}_i\right)\right]^2}{\displaystyle\sum_{j=1}^{53}\left(1\Big/\left[\mathrm{SE}\left(\hat{\beta}_j\right)\right]^2\right)}. \qquad (2)$$

Assuming the $\hat{\beta}_i$'s are independent, the natural estimate

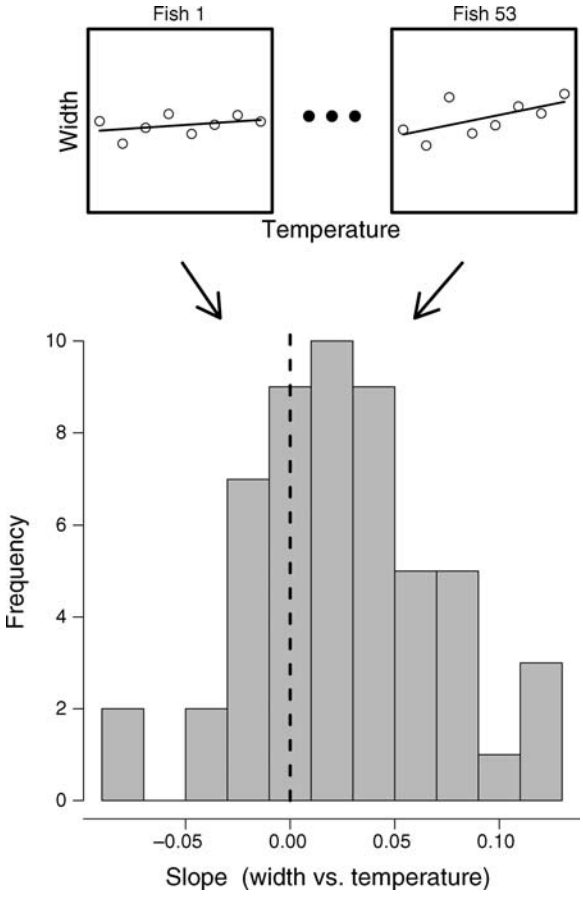

FIG. 3. For 53 Lost River suckers collected in 1995 (Terwilliger et al. 2003), a histogram of the slopes of regressions of increment-width residuals vs. lake temperature, done separately for each of the 53 fish, as shown in the top panel and in Fig. 2c. The vertical dashed line in the histogram indicates zero slope. The average of the 53 slopes, weighted by the inverse of their estimated variances, is 0.0218 μm/°C.

of the standard error of the weighted average is calculated as follows:

$$\mathrm{SE}\left(\hat{\bar{\beta}}\right) = \sqrt{\frac{\displaystyle\sum_{i=1}^{53} w_i\left(\hat{\beta}_i - \hat{\bar{\beta}}\right)^2}{53-1}}. \qquad (3)$$

If, as is usually assumed, each $\hat{\beta}_i$ has a normal distribution, we can base one-sample statistical inference on the fact that $(\hat{\bar{\beta}} - \beta)/\mathrm{SE}(\hat{\bar{\beta}})$ has a t distribution with 52 degrees of freedom. The estimates $\hat{\bar{\beta}}$ and $\mathrm{SE}(\hat{\bar{\beta}})$ can be obtained from most regression packages by fitting an intercept-only model of the 53 $\hat{\beta}$'s, with weights equal to the reciprocals of the squared standard errors from the 53 fish-specific fits.

This approach assumes only that there is some underlying β relating increment width to temperature for this population of fish, for which $\hat{\beta}_1, \ldots, \hat{\beta}_{53}$ are independent estimates. Whether β is a fixed effect for all

TABLE 2. Estimates of the regression coefficient relating increment width to temperature for the 53 Lost River suckers, based on a variety of statistical approaches.

General approach and temperature effect	Assumed correlation structure	Estimate	SE
Linear mixed effects			
Fixed	none	0.0224	0.0044
	AR(1)	0.0195	0.0071
	MA(1)	0.0219	0.0054
Random	none	0.0235	0.0057
	AR(1)	0.0195	0.0071
	MA(1)	0.0219	0.0054
Individual based		0.0218	0.0059

Note: All of the LME results were obtained using restricted maximum likelihood (REML).

fish, or the mean of fish-specific random effects (cf. Eq. 1), is irrelevant, as is the correlation structure of the errors within individual fish.

Comparison of approaches

Table 2 shows estimates obtained using variants of each of these methods. The individual-based method estimates that a one-degree increase in temperature is associated with a 0.0218-μm increase in otolith increment width (SE = 0.0059). The estimates based on the linear mixed-effects models range from 0.0195 to 0.0235 μm/degree, with the choice of correlation structure having a large influence on the result. The standard errors from the mixed-effects models are also quite variable, with the largest (0.0071) being 61% larger than the smallest (0.0044).

Of course, one cannot determine by inspection which of the estimates in Table 2 is "best." To explore the validity of the different approaches in this example, I simulated data by (1) generating intercepts and slopes of 53 fish-specific regressions of increment width vs. temperature from a bivariate normal distribution based on the empirical data, and (2) adding time series of errors of the appropriate length to each regression line, with correlation structure dictated by the distributions of residuals in the original data. Details of these simulations are given in Appendix B.

I then applied the various analysis methods to the simulated data; recorded the slope estimates and their standard errors; and constructed confidence intervals for the population mean slope, noting whether or not each interval included the true slope (used in simulating the data). Figure 4 shows the results of these simulations.

The individual-based method (analysis 2) yields slope estimates that are, on average, very close to the true value, and the confidence intervals have the expected 95% coverage. The LME models treating the temperature effect as random (labeled RE in Fig. 4) are reasonably accurate, with confidence-interval coverage close to the nominal level. The models that treat temperature as a fixed effect (FE) behave more erratically. The confidence intervals have smaller than the nominal coverage, because the estimated standard errors are too small.

It is not surprising that the LME models treating temperature as a random effect do better than those treating temperature as a fixed effect, since the simulation algorithm is essentially a random-effects model (see Appendix B). Of course, in analyses of real data, the underlying model is not known, and the choice of how to model the temperature effect may be quite subjective.

As mentioned earlier, the choice of correlation structure in LME modeling can have a large effect on the results (Fig. 4). In addition, for certain models, different estimation algorithms (REML or ML) may yield quite different results—see, for example, the results for fixed-temperature-effect models assuming no within-fish correlation of errors (labeled FE, NONE).

This single set of simulations does not prove that the performance of the individual-based method is always as good as or better than those of the LME modeling approaches. But the simplicity of the assumptions behind the individual-based approach—that the fish-specific regression coefficients are independent, normally distributed random variables—suggests that this approach is likely to work well under a variety of possible scenarios of data generation.

The results of the LME approaches, on the other hand, are potentially quite sensitive to the choices that

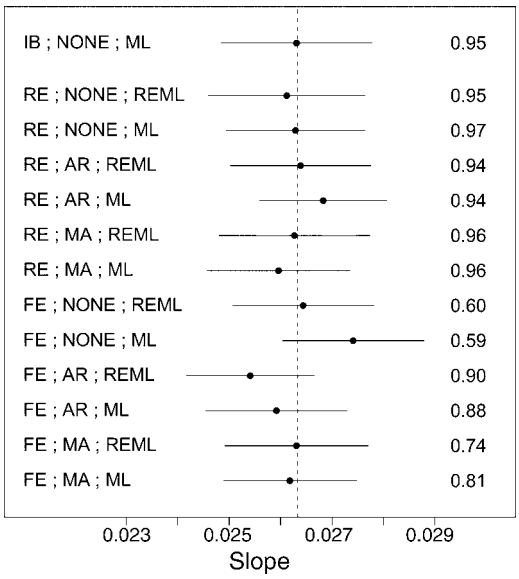

FIG. 4. Estimates of the regression coefficient (in μm/°C) for temperature vs. increment width from analyses of data sets simulated as described in Appendix B. Lines show 95% CI based on the empirical variance of 200 simulated slope estimates. The labels on the left give the analysis method (IB, individual based; RE, temperature as a random effect; FE, temperature as a fixed effect); the correlation structure assumed for the errors (none, first-order autoregressive, or moving average); and the estimation algorithm (maximum likelihood or restricted maximum likelihood). The numbers on the right give the proportions of 200 simulated confidence intervals that included the true mean slope.

need to be made in implementing the analyses, mostly pertaining to the modeling of the correlation of observations within units. Some statisticians may be able to use experience, intuition and subtle techniques of data exploration to guide the decisions that must be made in fitting LME models (e.g., see Venables and Ripley 2002: 272–279), but many other users, including this one, will be overwhelmed by the technical detail that must be mastered in order to use these methods appropriately and effectively.

This is not to say that linear mixed-effects models are not useful; they are invaluable when there is specific interest in the nature of the variation of responses made within units, and the way that the within-unit variability interacts with the between-unit variability to determine the overall pattern of responses. If the questions of interest focus on the larger units, however, the analysis can often be simplified by summarizing, rather than modeling, the multiple observations made within units—in this example, using a least-squares regression estimate, instead of a time-series model, to summarize each fish's data.

Other Issues

There is a variety of other ways that ecologists might simplify and strengthen their analyses. One is to use more graphs. In my opinion, one should be skeptical of a statistical conclusion that is unaccompanied by some graphical demonstration of the trend or difference being tested. Graphical exploration is a good first step of any analysis, with statistics often playing a secondary, confirmatory role. Nonetheless, complicated statistical analyses are often presented without graphs. For example, a long discussion of the effects of forest practices on peak flows in watersheds in the western Cascades of Oregon (Thomas and Megahan 1998, Beschta et al. 2000, Jones 2000) includes no examples of the hydrographs upon which the analyses are based.

As in the examples explored here, complexity is often introduced when analyses are based on multiple measurements made on a relatively small number of experimental or observational units. Substantial effort may be devoted to the modeling of the measurements within units, when often all that is needed is a single summary of those measurements at the level of the experimental unit. O'Brien and Shampo (1988) describe a study of time series of blood flow in eight subjects in which there is no attempt, or need, to model the serial correlation of measurements within subjects: "We believe that, in this instance, the use of a graphic display aided by *t* tests appropriately conveyed the pertinent information obtained by the investigators and that the use of more sophisticated statistical analysis would not have been helpful."

Conclusions

Some statistical methods are unavoidably complex, e.g., survival analysis, mark-recapture analyses, many applications of Bayesian inference, and spatial statistics. Presentations of such analyses will necessarily be quite technical and involved. But, in more "run-of-the-mill" analyses, simplicity is a worthy goal to strive for. Simple analyses offer fewer chances for mistakes; they are easier to explain and understand; they require fewer arbitrary choices by the analyst; and they are likely to lead to the same conclusions when applied by different analysts to the same data.

Of course, statistical rigor should not be sacrificed in the quest for directness; there are many examples of simple analyses that are misleading or incorrect (e.g., see Hurlbert 1984). Identifying the crux of a statistical problem may require considerable statistical training, experience and intuition, even if the ultimate solution involves a commonplace technique.

If statistical analyses were more transparent, it would be easier for readers and reviewers to judge the validity and strength of a study's conclusions. At a minimum, the methodology should be understandable to a professional statistician; it should be explained in enough detail that the statistician could reproduce the analysis; it should be supported by graphical or descriptive summaries of the data; and it should be simple and direct enough that two analysts using the same approach are likely to reach the same conclusions. In my opinion, the quality of the ecological literature would improve if these criteria were more often met.

Acknowledgments

I am grateful to two anonymous reviewers for detailed and constructive criticisms of an earlier version of the manuscript and to Michael Schlax for helpful discussions.

Literature Cited

Beschta, R. L., M. R. Pyles, A. E. Skaugset, and C. G. Surfleet. 2000. Peakflow responses to forest practices in the western Cascades of Oregon, USA. Journal of Hydrology 233:102–120.

Cleveland, W. S. 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. American Statistician 35:54.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger. 2002. Analysis of longitudinal data. Second edition. Oxford University Press, New York, New York, USA.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54:187–211.

Jones, J. A. 2000. Hydrologic processes and peak discharge response to forest removal, regrowth, and roads in 10 small experimental basins, western Cascades, Oregon. Water Resources Research 36:2621–2642.

Kuehl, R. O. 2002. Design of experiments: statistical principles of research design and analysis. Second edition. Brooks/Cole, Pacific Grove, California, USA.

Murtaugh, P. A. 1989. Size and species composition of zooplankton in experimental ponds with and without fishes. Journal of Freshwater Ecology 5:27–38.

Murtaugh, P. A. 2000. Paired intervention analysis in ecology. Journal of Agricultural, Biological, and Environmental Statistics 5:280–292.

O'Brien, P. C., and M. A. Shampo. 1988. Statistical considerations for performing multiple tests in a single experiment. 3.

Repeated measures over time. Mayo Clinic Proceedings 63: 918–920.
Pinheiro, J. C., and D. M. Bates. 2000. Mixed-effects models in S and S-PLUS. Springer-Verlag, New York, New York, USA.
R Development Core Team. 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Schluter, D. 1994, Experimental evidence that competition promotes divergence in adaptive radiation. Science 266:798–801.
Stewart-Oaten, A. 1996, Problems in the analysis of environmental monitoring data. Pages 109–131 *in* R. J. Schmitt and C. W. Osenberg, editors. Detection of ecological impacts: conceptual issues and application in coastal marine habitats. Academic Press, San Diego, California, USA.
Terwilliger, M. R., D. F. Markle, and J. Kann. 2003. Associations between water quality and daily growth of juvenile shortnose and Lost River suckers in Upper Klamath Lake, Oregon. Transactions of the American Fisheries Society 132:691–709.
Thomas, R. B., and W. F. Megahan. 1998, Peak flow responses to clear-cutting and roads in small and large basins, western Cascades, Oregon: a second opinion. Water Resources Research 34:3393–3403.
Venables, W. N., and B. D. Ripley. 2002, Modern applied statistics with S. Fourth edition. Springer-Verlag, New York, New York, USA.

APPENDIX A

Equivalence of the *F* test and *t* test in Example 1 (*Ecological Archives* E088-003-A1).

APPENDIX B

A description of how data were simulated for the evaluation of the different analysis techniques for Example 2 (*Ecological Archives* E088-003-A2).