

PSEUDOREPLICATION ISSUES VERSUS HYPOTHESIS TESTING AND FIELD STUDY DESIGNS

Alternative study designs and statistical analyses help prevent data misinterpretation

By ROY IRWIN AND LYNETTE STEVENS

Editor's note: *Pseudoreplication issues are complex, and space constraints allow only an introduction here. A more detailed recap and a related summary of impacts considered de minimis¹ (small enough to be trivial) is available through e-mail from Roy Irwin, "roy_irwin@nps.gov".*

PSEUDOREPLICATION HAS become a popular buzzword that has attracted considerable interest, controversy, and confusion. The debates over pseudoreplication began with Hurlbert's introduction and definition of the term pseudoreplication as:

The use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent (Hurlbert 1984).

Since 1984, many papers have attempted to refute, better explain, and expand on the issue of pseudoreplication. Pseudoreplication occurs when classical hypothesis-testing treatments are not technically replicated or statistically independent. Pseudoreplication often involves (but is not limited to) situations where investigators extrapolate site-specific statistical inferences beyond the situation that was studied. Pseudoreplication does not describe just a particular type of experimen-

tal design, but rather a particular category of misinterpretations or incorrect analyses.

Although many NPS studies involve routine inventory and monitoring rather than experiments to document effects of various stresses, staff doing inventory and monitoring studies should be aware of pseudoreplication issues. In many cases, others will eventually compare past and present data in an attempt to get insight as to whether or not a trend is developing or whether or not some impact (a treatment) is causing resource deterioration. Therefore, inventory and monitoring studies should be designed to maximize their utility for future trend or injury analyses.

Pseudoreplication does not describe just a particular type of experimental design, but rather a particular category of misinterpretations or incorrect analyses

EXAMPLE

A common example of pseudoreplication occurs when repeated observations of a subject are substituted for replicated applications of a treatment on different subjects. In this situation, the sample design calls for taking measurements over time, but uses only one control and one treated site (subject); data are not spatially replicated. For example, consider a common before-and-after study to determine effects on aquatic biota from some point source effluent entering a stream (e.g., a power plant on a river).

The study design consists of two sample points: one above the input and one below. Although several samples may be collected above, and several below, these are not true "replicates" for purposes of hypothesis testing coupled with inductive (from specific case to general case) inference, since there is only one treatment (the power plant effluent) and one experimental unit (the specific river). Due to lack of true replication, inductive statistics should not be applied. In other words, you cannot use the results of this study to generalize about any other power plants or other stream systems, or even to conclude that the power plant caused the difference seen in this one situation.

A key point to keep in mind is that Hurlbert's original definition concerned pseudoreplication with respect to testing effects of treatments. By common (mis)usage, some have also used the term pseudoreplication more broadly to include such things as no replication or inappropriate replication, even in the absence of an effort to examine treatment effects (cause and effects) through the use of inductive statistical inferences. Examples would include the following: (1) taking three sediment samples from one area, thoroughly mixing the three in a pan, putting portions of the mixed sample in three separate jars, and then calling the

three jars three "replicate" samples; or (2) taking three samples so close together in time or space that they are really more like one sample than three samples.

In routine monitoring, these examples of questionable replication, in the absence of treatment effects testing, may be considered unwise or inappropriate, but would not be considered pseudoreplication under Hurlbert's original definition. Although often done, criticizing a data set as "pseudoreplicated" is usually inappropriate unless statistical inferences are made for cause and effect.

¹ The phrase *de minimis* is an abbreviated form of the Latin phrase "de minimis non curat lex," which translates to "the law cares not for small matters;" in risk assessment, *de minimis* impacts are those that are so small (and not related to special resources such as endangered species) that one can disregard them.

PSEUDOREPLICATION ISSUES IN ECOLOGY

Due in part to large amounts of natural variance, lack of baseline ecological data, and lack of adequate funds for complete replication of studies (or treatments), occurrence of pseudoreplication is especially

cedures. **Illustration**—If you measured the length of 97 plants in a quadrat, report one number, (the average or median, for example). Avoid the temptation to say $N=97$; instead, say $N=1$.

Red flag #4—Measurements on the effects of a single point source on a river, up- and downstream. Limiting the statistical inference to that unique location above and below is not pseudoreplication.

Note: It is safest not to expand your inferences beyond that one site and situation and not to pretend that your significant levels or various statistical inferences prove the cause of any differences noted.

STATISTICS APPLIED TO THE SCIENTIFIC METHOD AND RISK ASSESSMENT

The ideal way to build sci-

entific knowledge is to use the scientific method in true experiments. A typically recommended scenario would be to combine genuine replication with random assignment of treatments to experimental units or probabilistic sampling from a study area (personal communication, Lyman L. McDonald, West Inc., 1995). However, while genuine replication is a powerful tool that should be used when possible, the scale of ecological research should not be dictated by statistical constraints (Hargrove and Pickering 1992).

In the absence of truly random samples, convincing evidence of an effect requires the effect to be demonstrated consistently at different times in different places (Meyer et al 1994). Consistent effects of incidents comprise a non-statistical type of inference. Such inferences are deductive (general case to specific case) or nonstatistical (Meyer et al. 1994). Although difficulty in replicating large-scale field manipulations makes quantifying

high in ecological research. Pseudoreplication is not inevitable if an experiment lacks treatment replications, but occurs only if the researcher misleads the reader by applying inappropriate statistical analyses or misstating the strength of the evidence obtained (Hargrove and Pickering 1992).

However, lack of true replication should not be portrayed as an evil to be avoided in all situations. Often pseudoreplicated (or unreplicated) studies cannot be avoided in disciplines such as medicine, regional ecology, observational field ecology, and astronomy, but the state of "scientific" knowledge or site-specific or issue-specific understanding of issues still slowly progresses to new heights, mostly through weight-of-evidence approaches. Just as astronomers cannot directly manipulate stars, and therefore are prone to pseudoreplication, regional ecology research is difficult or often impossible without pseudoreplication (Hargrove and Pickering 1992). While population studies of animals with large home ranges may not be appropriate for experimentation, sample surveys and demographic studies may still be used to assess population effects (Skalski and Robson 1992).

The following are red flags for potential pseudoreplication problems (Meyer et al 1994):

Red flag #1—Use of more than one data point from given random or systematic plots or locations. **Solution**—Use one datum per sampling unit in statistical pro-

Notes: Many contaminants and biology effects data sets are not normally distributed, so it is often preferable to use nonparametric methods, to dispense with means and variances altogether, and to utilize alternative descriptive measures of central value and variability for skewed data, such as the median and interquartile ranges—IQR (Heisel 1990). In epidemiology work, a common practice is to compare only three groups: those with clearly high doses, those with clearly low (or no) doses, and an intermediate group. Even random selection of study sites from assessment and reference areas is "subsampling" or pseudoreplication if statistical conclusions are extrapolated beyond the assessment and reference areas (Meyer et al 1994).

Red flag #2—Use of the same plots or locations over time (OK for determining what happens to that one plot but not for larger universes).

Red flag #3—Multiple observations on same animal (OK for determining what happens to that one animal but not for determining what happens to populations).

The problem is often triggered when the investigator uses a classical null hypothesis testing analysis and overstates conclusions related to the causes of the differences

Continued on page 30

cause and effect relationships difficult, this loss of statistical inference to pseudoreplication may be offset by carefully developing ecological inferences (Hargrove and Pickering 1992).

the statistical conclusions of a manipulative experiment extend to the protocol by which the study was conducted while the statistical conclusions of an observational study are limited to the specific assess-

and *Ecological Risk Assessment*. A synopsis of the information is presented here with permission of Glenn Suter):

"In ecological epidemiology there is no random assignment of populations or communities to treatments and treatments are almost never replicated so we cannot use statistics to test the hypothesis that populations or communities treated with a pollutant are different than those that are not.

There is no truly random assignment of treatments. The investigator cannot randomly assign some reaches to be treated with an effluent and others not to be treated.

In other words, while hypothesis testing requires random assignment of treatments, the investigator has typically had no role in where the effluent pipe was placed.

In field studies there is often only one treatment (an effluent, for example) rather than replicated treatments. Multiple biological samples (for example, benthic macroinvertebrates) are often taken above and below the effluent. However, the downstream samples are taken from one community affected by a single effluent and the samples are pseudoreplicates from that one treatment. In other words, samples from above and below a discharge pipe are not true replicates, they are pseudoreplicates since there is only one treatment.

The question often arises: can't we use hypothesis testing if we do it right? There have been heroic efforts to do so (for example, the Stewart-Oaten, BACI design which nevertheless does not totally solve the problem).

Since there is an inherent bias in favor of the null hypothesis, hypothesis testing places disproportionate burdens on environmental protection. Those who would protect the environment are required to prove with 95% confidence that effects are occurring. This bias is defensible in pure science but indefensible in risk assessment; it rewards polluters who perform poor studies with few replicates and high variances due to sloppy techniques. Hypothesis testing provides less protection for organisms less abundant or more difficult to sample.

Pseudoreplication can be avoided by applying truly replicated treatments or by restricting the generality and comprehensiveness of one's conclusions.

Ecological (deductive, weight-of-evidence) but not statistical (inductive) inferences can be made even when treatments are not replicated (personal communication, Lyman L. McDonald, West Inc., 1995). For example, normic statements (i.e., statements of what usually or normally happens that are generated by the collective outcomes of repeated experiments) can be the result of pseudorelicated experiments. Although these statements are not universal, probability based, or predictive, they represent generalizations with exceptions. Even so, the information content of normic statements is high in terms of explanatory power (Hargrove and Pickering 1992).

Some of the problems related to pseudoreplication in field studies arise because the investigator is conducting an observational study rather than a manipulative experiment. The problem is often triggered when the investigator then uses a classical null hypothesis² testing analysis and overstates conclusions related to the causes of the differences. This has become common partly because most standard textbooks dealing with statistics and biological study design do not adequately distinguish between statistical conclusions drawn from manipulative experiments and statistical conclusions drawn from observational studies (Meyer et al 1994). The arithmetic analysis is often the same for both types of design, but

ment area, reference area(s), baseline conditions, and assessment period in the study (Meyer et al 1994).

HYPOTHESIS TESTING AND PSEUDOREPLICATION

Pseudoreplication problems are often partly the result of inappropriately trying to force non-replicated data into classical null hypothesis testing molds. Descriptive statistics, predictive methods, or various other observational data analysis methods are often more appropriate than classical null hypothesis testing schemes for environmental observational studies, injury assessments, many before-and-after (BACI) applications, upstream-downstream studies, and ecological risk assessments.

At least some hypothesis testing assumptions are typically violated in field studies. A thorough discussion of the pitfalls of using hypothesis testing in field study applications would require a separate article, but the following notes from Suter regarding hypothesis testing versus risk assessment and field studies are helpful in understanding pseudoreplication issues. The following points were presented briefly in Suter's *Ecological Risk Assessment* Text book (Suter 1993) and expanded in a platform session paper presented at the National SETAC meeting in Denver, November 3, 1994: entitled "The abuse of hypothesis testing statistics in ecological risk assessment." A paper of the same title and basic content was in press as of January 1996, in *Human*

² The null hypothesis is the hypothesis that an observed difference (as between the means of two samples) is due to chance alone and not due to a systematic cause.

Polluters love hypothesis testing since it can be done with poor data.... They can then fail to reject the null hypothesis and the environment is not cleaned up.... Our real problem is often defining real significance.... Often we should use descriptive rather than experimental statistics."

SOME SOLUTIONS

Two study sites in a single area, for example, just up- and downstream of a discharge, can properly be sampled and compared using descriptive statistics relating to magnitudes, variances, and trends. Some would say they could also be compared using a classical null hypothesis testing scheme, while others would say a null hypothesis testing scheme should not be used. Both would agree that no matter which statistical methods are used, if a difference is shown, the investigator only knows that they are "different," has not proved why they are different, and should be careful not to generalize the results to other sites, times, or conditions that were not studied. Although field researchers have sometimes determined that the samples are "different" using hypothesis testing, conclusions as to why they are different often cannot be drawn (Suter 1993).

Some experts say the results (that the samples are different) of hypothesis testing at one site could be used as one of several clues making up a weight-of-evidence argument related to effects at that one site. Other experts would argue that it is better to use descriptive statistics to suggest that the samples are "different" and thereby avoid any hint of an incorrect conclusion that cause and effect has been "proved" at any given significance level.

A key point to keep in mind is the importance of properly limiting inductive inferences or conclusions. Pseudoreplication can be avoided by applying truly replicated treatments or by restricting the generality of one's conclusions (i.e., not overstating results) (Dixon and Garrett 1994). Some would argue that restricting the generality of one's conclusions might sometimes involve stating that you do not know why the samples are different.

Instead of hypothesis testing, using descriptive statistics and a weight-of-evidence approach to link potential relationships is often better (Suter 1993). Field situations are often uncontrolled, unreplicated, and typified by so many confounding variables that cause and effect is difficult to establish. The weight-of-evidence approach (which often includes statistical data from both field and lab sources) is often safer in ecological, risk-injury assessment, and contaminants field work (Suter 1993, Chapman 1995).

It is sometimes acceptable to set up the analysis in terms of tests of bioequivalence in the following manner: Assume the treatments will result in a difference in bioequivalence, including variation up to de minimis amounts of acceptable natural variation. The investigator determines the level at which there can be a change without exceeding bioequivalence thresholds (for example, a percentage change in an endpoint such as biomass or number of taxa). If the effect does not exceed a certain percentage previously chosen by the investigator as a trivial or de minimis change, the change has not exceeded a bioequivalence threshold. Such an approach requires the investigator to deal with issues of natural variation and confidence.

CONCLUSIONS

Understanding the intricacies of pseudoreplication and hypothesis testing versus field study design issues is not an easy task. The importance of proper determinations of the interrelationship between study objectives, study designs, data analyses, and statistical inferences in field investigations cannot be stressed enough (Skalski and Robson 1992). Those lacking expertise in statistics, may find it wise to first consult a statistician familiar with pseudoreplication and the study design issues discussed herein (not just any handy statistician or book). It is also wise to develop a written (and defensible) statistical design plan prior to beginning the study.

PS

REFERENCES CITED

- Chapman, P.M. 1995. Extrapolating laboratory toxicity results to the field. *Environmental Toxicology and Chemistry* **14**:927-930.
- Dixon, M.D. and K.A. Garrett. 1994. Statistical issues for field experiments. Pages 439-450 *In* R.J. Kendall and T.E. Lacher, Jr., editors. *Wildlife Toxicology and Population Modeling: Integrated Studies of Agroecosystems*. SETAC Special Publication Series. Lewis Publishers, Boca Raton, FL.
- Hargrove, W.W. and J. Pickering. 1992. Pseudoreplication: a *sine qua non* for regional ecology. *Landscape Ecology* **6**(4): 251-258.
- Heisel, D.R. 1990. Less than obvious. *Environmental Science and Technology* **24**(12):1767-1774.
- Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **53**:187-211.
- Meyer, J.S., S.L. Hill, A.M. Boelter, J.C.A. Marr, A.M. Farag, R.K. MacRae, J.A. Hansen, M.J. Szumski, T.L. Parish, H.L. Bergman, L. McDonald, G. Johnson, D. Strickland, T. Dean, and R. Rowe. 1994. Proposed guidance document for determination of injury to biological resources resulting from incidents involving oil, draft report. Submitted to (but not yet approved by) the National Oceanic and Atmospheric Administration Damage Assessment Regulations Team, Washington, D.C. by Harold Bergman, University of Wyoming, June, 1994. Permission to quote information granted to NPS from Project Team Leader Harold Bergman.
- Skalski, J.R. and D.S. Robson. 1992. *Techniques for wildlife investigations: design and analysis of capture data*. Academic Press, Inc. San Diego, CA., Reproduced with the permission of the publisher. 237 pp.
- Stewart-Oaten, A., W.W. Murdoch and K.R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? *Ecology* **67**(4):929-940.
- Suter, G.W. II. 1993. *Ecological Risk Assessment*. Lewis Publishers, Chelsea, MI. 538 pp.

Dr. Roy Irwin is a Senior Contaminants Specialist and aquatic biomonitoring coordinator and Lynette Stevens is an Assistant Contaminants Specialist. They are located at the Water Resources Division of the National Park Service, Fort Collins, Colorado, and can be reached by phone (970) 225-3520, fax (970) 225-9965, or e-mail at "roy_irwin@nps.gov" or "lynette@meecker.cnr.colostate.edu", respectively.