COMMENTARY

Common Statistical Mistakes in Entomology: Pseudoreplication

DALE W. SPURGEON

A prevalence of statistical problems in the ecological (and entomological) literature is well documented. Since 1986, in publications of the Ecological Society of America alone, at least five articles have addressed common statistical errors (Wang 1986, Day and Quinn 1989, Fowler 1990, Yoccoz 1991, Bennington and Thayne 1994). These errors continue to occur because of a paucity of simple and usable advice, and because statistical software typically provides output that appears legitimate when it is not. The most troubling errors reflect basic violations of experimental design and analysis and include pseudoreplication, analyses that are inconsistent with the experimental design, interactions that are misrepresented or ignored, and ineffective blocking or inappropriate inference space. These errors are common, and although they may not always lead to wrong conclusions, they often do.

This commentary is the first in a series of commentaries that will be published in *American Entomologist* to address common statistical mistakes in entomology, aligned with the four errors mentioned in the previous paragraph. This first commentary addresses pseudoreplication, with simple advice to facilitate its detection and avoidance.

A reasonable analysis provides evidence to guide or support interpretation of experimental results. It also reflects the physical layout of an experiment, which is necessary to construct meaningful tests of hypotheses. Pseudoreplication, whether it results from a lack of true replication or incorrect specification of the statistical model, is not consistent with a reasonable analysis. Pseudoreplication occurs when a sampling or subsampling unit is used as if it was a true replication, or experimental unit (EU), although pseudoreplication sometimes cannot be avoided. Examples include "areawide" experiments where more than one comparable EU is not available, or where true replication is precluded by the high cost of the treatments. However, these situations are uncommon.

To avoid pseudoreplication, one must recognize the EU to which a treatment is assigned. In a small-plot field study, the true sample size is the number of EUs (plots) being treated, regardless of the number of samples taken from each plot. In a laboratory study where a treatment is assigned to an environmental chamber, the chamber is the EU, regardless of the number of insects it contains. When groups of insects are assigned to treatments, as in a bioassay, the EU is the bioassay group treated with a particular material or dose, regardless of the number of insects within each group. In each of these examples, pseudoreplication exists because the responses of subjects within each plot, chamber, or dose are not independent. Responses of insects or plants within a group (plot, chamber, or application of a dose) will generally be more alike than responses of subjects from different groups receiving the same treatment. If treatments are assigned to individual subjects that are otherwise treated identically, and their responses are independent, then the individuals may be the sampling unit and the EU.

The distinction between the sampling unit and the EU is important because, in

an analysis of variance (ANOVA), the magnitude of a treatment effect is assessed by a ratio of variances. The variance is the average squared difference (mean square) for a group of responses from their overall mean. The total variance is the mean squared difference of all the EUs from the overall mean. In ANOVA, the total variance is *partitioned* among the sources of variation (treatment, block, etc.) to form *variance components*. The treatment variance is the mean squared difference between the treatment means and their overall mean. Estimates of the total variance and of variance components (e.g., treatment, block) are based on *inter*-plot (among EU) variation. Experimental error is estimated by subtracting the other variance components from the total variance, so it is the variation among responses of EUs in the absence of other sources of variation. In contrast, variances based on subsamples from within each plot are measures of *intra*-plot variation. When variance components of the model (treatment, block, error, etc.) are estimated based on subsamples, the estimates are entangled mixtures of inter-plot and intra-plot variances.

If the treatments have no effect, the variance among the treatments is no different than the variance among the EUs (error variance). If, however, the treatment means are different, the variation among treatments is greater than the error variance. The *F*-test assesses the probability that (treatment variance/error variance) >1. When pseudoreplication is used so that the error variance is a mixture of inter- and intra-plot variation, this ratio of variances is not meaningful. Because the number of subsamples is larger than the number of EUs, the df for the error variance is also inflated, which lowers the critical value of the *F*-statistic. This situation increases the likelihood of a type-I error (a treatment effect is declared where none exists). If the statistical model does not correctly identify the EU, then the *F*-test is nonsensical and invalid. This scenario occurs in the literature with alarming frequency.

Consider a randomized complete block design with one treatment effect (Trt) comprised of three levels (e.g., plant varieties). Each variety is randomly assigned to a plot within three blocks, representing a total of nine EUs (Fig. 1). Assume there are 10 observations per plot (subsamples), yielding a total of 90 observations. Subsampling is common because it provides a better estimate of the plot response compared with a single sample per plot. Analyzing this design using PROC GLIMMIX of SAS (SAS Institute 2012), where the responses are the means of the subsamples from each plot (such that each EU is represented by one observation) is straightforward:

```
proc glimmix;
class trt block;
model response=trt;
random block;
run;
```

Because we have nine EUs (total df = 9 – 1 = 8), three treatment levels (df = 2), and three blocks (df = 2), the residual variance with 4 df (8 – 2 – 2 = 4) is also the error variance for testing Trt. However, if this model is used to analyze the data including the subsamples, the results are strikingly different, because SAS interprets the subsamples as if they were EUs, and the residual variance is represented by 85 degrees of freedom. A valid analysis requires inclusion of an appropriate error term representing the EUs, which can be uniquely identified by the combination of treatment assignment and block (Trt*Block):

```
proc glimmix;
class trt block;
model response=trt;
random block trt*block;
run;
```

This program partitions the experimental error (Trt*Block, 4 df) from the residual (now with 81 df), and the results are identical to the analysis of plot means. If the numbers of subsamples per plot were not equal, or there were missing observations, the results of these two analyses would not be identical. In those cases, the analysis using the subsamples would be more informative because PROC GLIMMIX (or PROC MIXED) uses the intra-plot information.

PROC GLIMMIX (or PROC MIXED) will generally construct the appropriate *F*-test if the correct error term is provided, with one caveat. If any estimate of covariance of the random effects is negative, SAS truncates the estimate to zero, drops the term from the model, and uses the incorrect error term. This is remedied by adding the "nobound" option to the PROC GLIMMIX (or PROC MIXED) statement. In contrast, PROC GLM does not construct the appropriate tests. A valid analysis including the subsamples in PROC GLM must associate the treatment effect (h=) with the appropriate error term (e=) through the "test" statement:

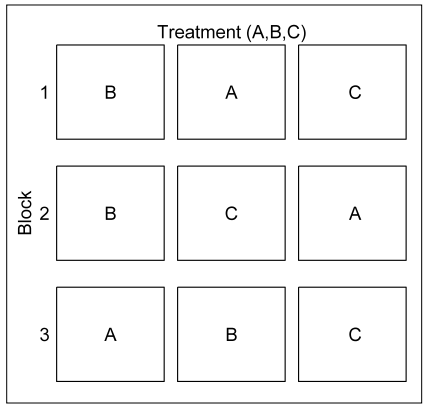

Fig. 1. Randomized complete block design with three levels of one treatment and three blocks, where boxes represent the experimental units (plots, chambers, doses, etc.) for Treatment.

```
proc glm;
class trt block;
model response=trt block
  trt*block;
random block;
test h=trt e=trt*block;
run;
```

This program outputs *F*-tests using residual as the error term (Trt, df = 2, 81; Block, df = 2, 81; Trt*Block, df = 4, 81), which should be ignored. Also, in this design no error terms exist for valid tests of Block or Trt*Block, so these tests are completely meaningless. The correct information is provided in the table "Tests of Hypotheses Using the Type III MS for Trt*Block as an Error Term," and it reports an *F*-test with the correct df (Trt, df = 2, 4).

But what if the design is completely randomized with no blocking? In that case, the combination of treatment and replication represents the plot, or EU (Fig. 2). In PROC MIXED (or PROC GLIMMIX), the Trt*Rep term represents the EU, but Rep is not included in the model

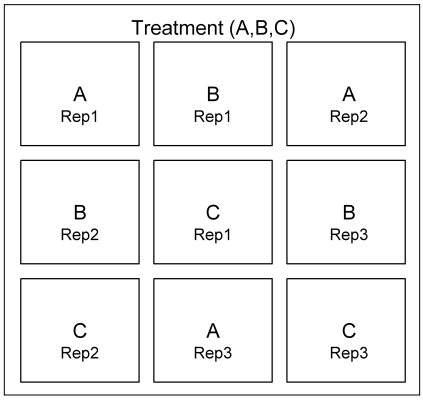

Fig. 2. Completely randomized design with three levels of one treatment, each level replicated three times, where boxes represent the experimental units (plots, chambers, doses, etc.) for Treatment.

because it does not represent a source of variation:

```
proc glimmix;
class trt rep;
model response=trt;
random trt*rep;
run;
```

This program properly tests the treatment effect (Trt, df = 2, 6) using the Trt*Rep interaction (df = 6) as error. If the random Trt*Rep term is excluded from the model, the *F*-test of Trt uses the residual as experimental error (df = 2, 87), which is meaningless. In contrast, PROC GLM requires the Trt*Rep term in the "model" statement or it cannot be used in the "test" statement:

```
proc glm;
class trt rep;
model response=trt trt*rep;
test h=trt e=trt*rep;
run;
```

Regardless of its inclusion, the *F*-test of Trt*Rep is not valid and should be ignored.

In more complex designs, the EU is still uniquely identified by its assignment to treatment combination and block (or rep). Consider where a second treatment effect (Trt2; irrigation with two levels) is added to the previous randomized block design, and each combination of Trt1 and Trt2 is assigned to a plot. If there are 10 subsamples per plot, the design features six plots in each of three blocks for a total of 18 EUs and 180 observations. An appropriate model would identify the EU (Trt1*Trt2*block) as the error for testing treatment effects:

```
proc glimmix;
class trt1 trt2 block;
model response=trt1 trt2
  trt1*trt2;
random block trt1*trt2*block;
run;
```

However, if Trt2 (irrigation) is instead applied to half of each plot, the experiment becomes a split-plot design with two sizes of EU (Fig. 3). The whole plot (Trt1*Block) represents the error term for Trt1, and the subplot (Trt1*Trt2*Block) represents the error term for Trt2 and the Trt1*Trt2 interaction. Therefore, the experiment includes nine whole plots as in the initial design, 18 subplots, and 180 observations. A valid analysis must provide error terms for both Trt1 and Trt2:

```
proc glimmix;
class trt1 trt2 block;
model response=trt1 trt2
  trt1*trt2;
random block trt1*block
  trt1*trt2*block;
run;
```

This program yields appropriate *F*-tests for Trt1 (df = 2, 4), Trt2 (df = 1, 6), and their interaction (Trt1*Trt2, df = 2, 6). The Trt1*Block interaction (df = 4) is used as error for testing Trt1, and the Trt1*Trt2*Block interaction (df = 6) is used as error for testing Trt2 and the Trt1*Trt2 interaction. Exclusion of either of the Trt1*Block or Trt1*Trt2*Block terms results in invalid *F*-tests, illustrating the importance of using a statistical model that is consistent with the physical design structure of the experiment.

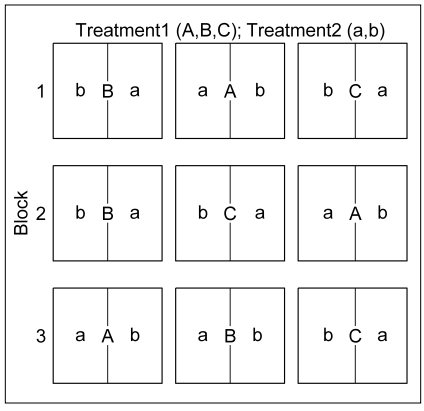

Fig. 3. Split-plot design with three levels of Treatment 1, two levels of Treatment 2, and three blocks, where boxes represent the experimental units (plots, chambers, doses, etc.) for Treatment 1, and half-boxes represent the experimental units for Treatment 2.

In summary, analyses that lack replication or that use inappropriate error terms are commonplace in the literature. These errors are easy to recognize because the model (unadjusted) error df for a given test cannot be larger than the number of EUs. If the analyses are conducted using PROC GLIMMIX or PROC MIXED, the analyst should verify that the error df are appropriate before applying the small-sample df correction (e.g., ddfm=kr) to the model statement. Because maximum-likelihood estimates are biased for small samples, a df correction should be used routinely. Although these corrected df are often non-integer, they will be of similar magnitude to the unadjusted df. These simple guidelines should allow authors, editors, reviewers, and readers to recognize and correct problems caused by pseudoreplication.

Acknowledgments

Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References Cited

Bennington, C.C., and W.V. Thayne. 1994. Use and misuse of mixed model analysis of variance in ecological studies. Ecology 75: 717–722.

Day, R.W., and G.P. Quinn. 1989. Comparisons of treatments after an analysis of variance in ecology. Ecological Monographs 59: 433–463.

Fowler, N. 1990. The 10 most common statistical errors. Bulletin of the Ecological Society of America 71: 161–164.

SAS Institute. 2012. SAS release ed. 9.4. SAS Institute, Cary, NC.

Wang, D. 1986. Use of statistics in ecology. Bulletin of the Ecological Society of America 67: 10–12.

Yoccoz, N.G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72: 106–111.

Dale W. Spurgeon, USDA, ARS, Arid-Land Agricultural Research Center, 21881 N Cardon Lane, Maricopa, AZ. E-mail: dale.spurgeon@ars.usda.gov

DOI: 10.1093/ae/tmz003