COMMENTARY

Common Statistical Mistakes in Entomology: Models Inconsistent with the Experimental Design

DALE W. SPURGEON

The first article of this series (Spurgeon 2019) addressed the common problem of pseudoreplication, its consequences, and how to recognize or avoid it. That article also described how the total variance within the experimental data is partitioned among sources of variation (*variance components*) associated with the model effects to provide an appropriate estimate of experimental error for comparison with a treatment variance. This comparison (treatment variance/error variance) produces the *F*-statistic used to assess the effect of the treatment. Both the numerator and denominator of this ratio contain the same components (called *expected mean squares*, EMS), except the numerator also contains the contribution by the treatment. All of the common terms in the numerator and denominator cancel, so when the contribution by the treatment is negligible the ratio (*F*-statistic) is close to one. When an analysis is conducted using pseudoreplication, the *F*-test of a treatment effect is misspecified because the EMS of experimental error no longer matches the corresponding EMS of the treatment variance. Therefore, the *F*-statistic is meaningless. In addition to pseudoreplication, there are other ways to incorrectly specify an ANOVA model, and simple ways to avoid such misspecification.

The purpose of experimental design is to partition the overall variance into components that represent model effects to be tested, adjust or control for influential but uninteresting sources of variation, and provide appropriate error terms for hypothesis testing. These objectives are not met when the analytical model does not represent the design of the experiment. Use of inappropriate ANOVA models is common in the literature; examples include a blocking factor intentionally or unintentionally excluded from the ANOVA model; a multifactor design divided into multiple one-way ANOVAs or *t*-tests; and, less commonly, an analysis that does not appropriately accommodate a single global control in a multifactor experiment.

Consider situations for which a blocking factor (whether or not it is called "Block") and/or the corresponding Trt*Block interaction are intentionally excluded from the ANOVA model. This exclusion is sometimes based on the *p*-values associated with their respective *F*-tests. As was illustrated in the first commentary (Spurgeon 2019), such exclusion is inappropriate, because unless the treatments are replicated *within* each block (a *generalized* randomized block design) there are no valid tests of the Block or Trt*Block effects. These terms are only useful for improving the estimate of experimental error (partitioning out the effect of Block) or as a representation of experimental error (Trt*Block) for testing the effect of Trt. Consider a randomized block design with three levels of one treatment and three blocks (Fig. 1). If there is one observation per experimental unit (EU; plot, chamber, etc.), the appropriate statements using PROC MIXED (or PROC GLIMMIX) of SAS (SAS Institute 2012) are:

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

In this model, the residual mean square is also the experimental error, which is

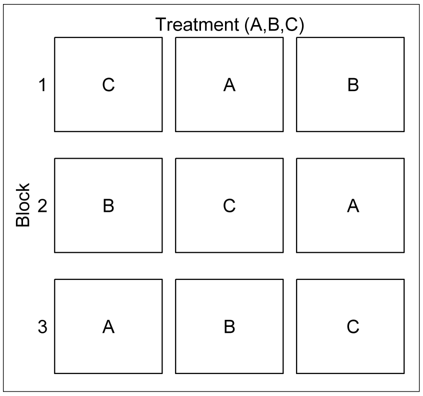


Fig. 1. Randomized complete block design with three levels of one treatment and three blocks.

not explicitly stated but is equivalent to the Trt*Block interaction. When Block is omitted from the model, its variance and df are absorbed by the residual mean square, which means the experimental error now contains a source of variation that is not represented in the treatment variance. The *F*-test obtained from these two variances is not meaningful unless the variation associated with Block is negligible, which cannot be tested.

If the EUs in Fig. 1 are each represented by 10 subsamples, the appropriate SAS statements must specify experimental error (Trt*Block), which is different from the residual:

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

Omission of Block and Trt*Block from this model not only fouls the estimate of experimental error, but results in pseudoreplication because the error term (Trt*Block) for testing Trt is eliminated.

It is not uncommon that a multifactor design is analyzed as multiple one-way ANOVAs. Consider a randomized block design with three blocks and two treatments, where each treatment is represented by two levels (Fig. 2). An appropriate analysis of this design, assuming subsampling within the EUs, is:

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

The term "Trt1*Trt2*Block" represents the EU and is the error term for testing effects of Trt1, Trt2, and Trt1*Trt2. If there is no subsampling, this three-way interaction is not explicitly stated because it is the same as the residual. A decision to analyze the effects of Trt1 and Trt2 in separate ANOVAs (or even worse, using *t*-tests) has several important consequences. First, if there is subsampling, identifying and constructing an appropriate error term for testing the effects of either treatment becomes difficult. Secondly, when one of the treatments is omitted from the model, its variance and df (and the variance and df of the interaction) do not simply disappear; instead, they are absorbed by other terms in the model. The obvious consequence is that variance components that should be common to both the treatment and error variances (numerator and denominator, respectively, of the *F*-statistic) no longer cancel, and it is not clear what, if anything, the *F*-test is testing. Finally, to interpret either tests of Trt1 or Trt2, one must assume the effects of the omitted treatment and the interaction between treatments are negligible, which are not reasonable assumptions.

Simple guidelines to indicate which terms should be included in the analysis are provided by Milliken and Johnson (1984). Similar advice can be found in Littell et al. (2006) and Stroup (2013), although their terminology is different from that of Milliken and Johnson (1984), who separate the experimental design into two components: the *design structure* and the *treatment structure*. The treatment structure is comprised of the factors imposed or selected by the researcher (generally, the fixed effects or treatments of explicit interest). The design structure is comprised of the physical entities that make up the experiment: the plots, replicates, blocks, strata, etc. In a typical mixed model, the treatment structure reflects the fixed effects, whereas the design structure reflects the random effects. In general, the treatment structure is represented in the ANOVA model by all of the fixed main effects (e.g., Trt1, Trt2) and all of their possible interactions (e.g., Trt1*Trt2). The design structure is represented by each blocking or stratifying effect (Block) and any of their interactions with treatment effects that are necessary as error terms for tests of hypotheses (e.g., Trt1*Trt2*Block). These latter interactions are random effects and represent the respective experimental units for each treatment effect.

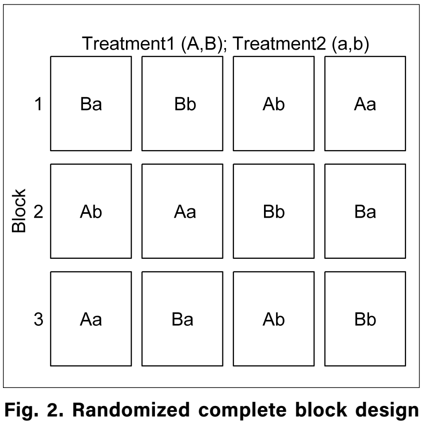
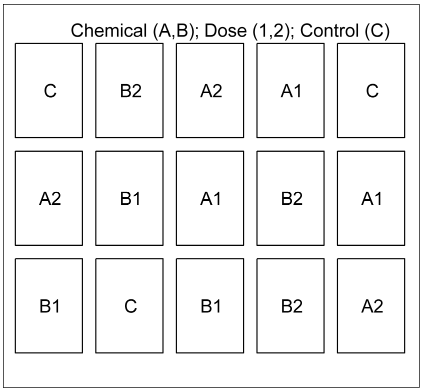


Fig. 2. Randomized complete block design with two treatments of two levels each and three blocks.

Finally, it is not uncommon to encounter multifactor experiments that include a single global control to which no treatment is applied. This seems especially common in bioassays and can be difficult for the reader to recognize. A single control used to represent all experimental repetitions is not very informative because the control is not replicated and no associated variance can be estimated. Occasionally, the control data are simply duplicated to allow analysis. This practice borders on data fabrication and still does not permit valid estimation of a control variance, because all of the repetitions of the control are identical. In this situation, the control represents a marginally useful indicator of the response in the absence of treatment, and its inclusion in an analysis is not appropriate.

In a multifactor experiment in which the control is appropriately replicated (i.e., a separate control is associated with each replication of the treatments), the typical effects model fails because the control has only one level. For example, consider a completely randomized design to examine virus transmission by a sucking insect in response to two chemical treatments (A, B) at two levels each (low and high recommended doses, 1 and 2, respectively). The EU is a single plant, the response is virus titer in the plant (Titer), and the experiment uses a global control (no chemical treatment). Each treatment combination and the control are replicated three times (Fig. 3). An *effects* model (the *effects* are Trt, Dose, and Trt*Dose) estimated by PROC MIXED uses the statements:

```
proc mixed;
class trt dose;
model titer=trt dose trt*dose;
run;
```



C	B2	A2	A1	C
A2	B1	A1	B2	A1
B1	C	B1	B2	A2

Fig. 3. Two-factor completely randomized factorial design replicated three times and with a global control.

Using this model, SAS reports *F*-statistics and *p*-values for the model effects, but they are not useful because at least some of the treatment (or dose) means cannot be estimated. The output from this model depends on how "Dose" is designated for "Control." If "Dose" is set to missing, then the control is excluded from the analysis. If the control is considered to represent "Dose" = 0, the means of the treatment levels (A, B, Control) and doses (0, 1, 2) cannot be estimated. Duplicating the control observations for each dose solves the computational problem, but misrepresents the experiment. Bergerud (1989) illustrates a procedure wherein the variance estimates obtained from multiple ANOVAs are used to construct contrasts to test the main effects (Trt, Dose) and interaction (Trt*Dose). That method is beyond the scope of this commentary but may be of interest to readers.

A less rigorous but still reasonable alternative to the effects model is a *means* model (Milliken and Johnson 1984), for which each treatment and dose *combination* is designated as a separate treatment. This can be accomplished by explicitly assigning a unique name to each treatment combination, or by constructing the combinations using the Trt*Dose interaction (e.g., `model titer=trt*dose;`. The respective outputs of these two approaches are identical. Although the main effects (Trt, Dose) are not tested in this analysis, means and standard errors corresponding to each treatment combination can be estimated, and meaningful comparisons can be made among them (adjusting for multiplicity, of course).

In summary, if one or more blocking effects (block, cohort, site, etc.) are physical parts of the experiment, they should be represented in the design structure irrespective of their estimated variances or *p*-values, which in most cases are uninformative. Because the *F*-statistic is a ratio of the variance of a given model effect to its error variance, it is important as much as possible to exclude other sources of variation from these terms. Every effect in an ANOVA model plays a role, whether it is a fixed effect of specific interest or a random effect that is part of the physical design. If effects are arbitrarily excluded from the model, the variances associated with those effects do not go away. Instead, they are embedded in the variances of other effects, and the objectives of experimental design are not realized. The simple guidelines herein should permit readers, reviewers, or editors to recognize analyses that are not consistent with reported experimental designs.

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

Bergerud, W. 1989. ANOVA: factorial designs with a separate control. Biometrics Information, Pamphlet 14. British Columbia Ministry of Forests Research Program. https://www2.gov.bc.ca/assets/gov/farming-natural-resources-and-industry/forestry/stewardship/biometrics-pamphlets/pamp14.pdf (Accessed 11 February 2019).

Littell, R.C., G.A. Milliken, W.W. Stroup, R.D. Wolfinger, and O. Schabenberger. 2006. SAS for mixed models, 2nd ed. SAS Institute, Cary, NC.

Milliken, G.A., and D.E. Johnson. 1984. Analysis of messy data, volume 1: designed experiments. Van Nostrand Reinhold, NY.

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

Spurgeon, D.W. 2019. Common statistical mistakes in entomology: pseudoreplication. American Entomologist 64: 16–18.

Stroup, W.W. 2013. Generalized linear mixed models: modern concepts, methods and applications. CRC Press, Boca Raton, FL.

Dale W. Spurgeon, USDA, ARS, Arid-Land Agricultural Research Center, 21881 N Cardon Lane, Maricopa, AZ. E-mail: dwspurg@gmail.com

DOI: 10.1093/ae/tmz022