

## PROC MIXED: Key concepts for appropriate mixed-models analysis

Dale W. Spurgeon and Sara Duke  
USDA-ARS, College Station, TX

In recent years, many of us have become aware that traditional ANOVA methods are not appropriate for most of the data we analyze. Consider the inferences drawn from an insect trapping study using fields, blocks, or plots as replicates. When these data are analyzed in SAS using PROC ANOVA or PROC GLM, the replicates are treated as fixed effects. Therefore, the inferences are narrow, and are only applicable to *the observed fields, blocks, or plots*. This is rarely our intent; we usually want our results to provide a broader inference that is relevant to the general population of fields, blocks, or plots similar to those we observed. This broad inference requires that we treat the replicates as random effects. Although a RANDOM statement is available in PROC GLM, the computational algorithm treats the replicates as fixed. Analyses in PROC GLM become even more problematic with the introduction of multiple levels of treatment replication (nesting or subsampling) or missing or unbalanced data. In these cases, PROC GLM does not construct appropriate tests of hypotheses by default, although many of us have not realized this was a problem, or did not know how this problem should be addressed.

A decade ago, a new SAS procedure, PROC MIXED, was developed to appropriately handle mixed-models (those with fixed and random effects). When properly used, PROC MIXED yields valid hypothesis tests, standard errors, and confidence intervals that correctly account for the random nature of the experimental units (EU's). Thus, the *inference space* is appropriately broadened to include fields, plots, or blocks beyond those we actually observed. Furthermore, PROC MIXED constructs appropriate hypothesis tests by default even in the presence of nesting, subsampling, serial correlations among observations, or unbalanced data – provided fixed and random effects are correctly specified in the MODEL and RANDOM statements. This last point is particularly relevant, because many authors using PROC MIXED recognize that the procedure constructs appropriate tests “automatically”, but are apparently unaware that misspecification of the model still results in the construction of meaningless or invalid hypotheses.

In the senior author's experience as a journal editor and reviewer, the use of PROC MIXED has noticeably increased even during the past year. Unfortunately, many authors are not using PROC MIXED appropriately, or they do not provide the information necessary to determine if the procedure is being used correctly. Herein we present five key concepts that should be addressed or grasped by the analyst in order to make a successful transition to PROC MIXED from the more traditional, fixed-effects approach represented by PROC GLM.

### **1. Change the approach to the model or hypotheses of interest.**

In traditional PROC GLM analysis, the interactions between treatment (fixed) factors and replication (random) factors were examined because these tests indicated whether the treatment effects were consistent among levels of the replication factors (fields, blocks, or plots). However, the mixed-models context assumes the observed replicates are a representative sample of all possible replicates. Consequently, we are explicitly interested in the *overall* treatment response across levels of field, block, or plot. In most instances, we would only be interested in, say, the treatment-by-field interaction if we suspected that one or more fields were *not*

representative of fields in general, and we had some reasonable insight to indicate this was the case. Otherwise, neither the MODEL nor RANDOM statements of PROC MIXED would contain a term for treatment-by-replicate unless this interaction was needed as an error term (denominator of an  $F$ -test) for one or more fixed effects.

## **2. Ensure that meaningful error terms are provided in the RANDOM statement.**

PROC MIXED provides valid hypothesis tests for the fixed effects only if the model is correctly specified. It is critical that the investigator understands the experimental design structure (the organization of the EU's) and how treatments are assigned to the EU's (3). Random effects (field, plot, or block) that identify sources of variability (variance components) are listed only in the RANDOM statement. Also in the RANDOM statement are any effects that are appropriate error terms for one or more hypotheses about fixed effects. The process of correctly defining these error term(s) can be confusing, however there are many proposed techniques for this purpose (1). The key point is that if the appropriate error term(s) are not specified, the analysis will be invalid.

## **3. Be aware of subsampling or hidden nesting and account for it.**

Authors frequently report collecting multiple samples per experimental unit but fail to account for such subsampling in analyses. For example, in a randomized complete block design with one fixed factor (treatment) and subsampling within plots, the block-by-treatment interaction (the statistical representation of the physical unit plot) is the appropriate denominator of the  $F$ -test for treatment. This is because plot represents the EU for treatment. Failure to include this interaction in the RANDOM statement causes PROC MIXED to use the (unstated) residual error as the denominator of the  $F$ -test for treatment. This approach treats the subsamples as if they were the EU's for treatment, and constitutes a form of pseudo-replication. Use of this error term usually results in artificially small  $P$ -values and standard errors because sample size is inflated. In such cases,  $F$ -tests with denominator degrees of freedom that are close to the total number of samples indicate that the wrong error term was used. In addition, nesting of EU's must be recognized. For instance, in a study with sites within fields and subsamples within sites, sites are actually nested within fields. This is because site represents a smaller EU within a larger EU (field). Thus, site-within-field [site(field)] should be included in the RANDOM statement. Also, tests of any treatment factors for which site is the appropriate experimental unit will require that the treatment-by-site-within-field interaction [treatment\*site(field)] be included in the RANDOM statement. Failure to account for such nesting will again result in invalid  $F$ -tests and standard errors.

## **4. Use a correction for the degrees of freedom.**

Because PROC MIXED solves the sets of linear equations using the Restricted Maximum Likelihood rather than Least Squares, the degrees of freedom corresponding to hypothesis tests must be approximated. If the experimental design is unbalanced, there are missing data, or if the design contains nesting, subsampling, or repeated measures, it is critical to specify the method for estimating the degrees of freedom. For most purposes the Kenward-Rogers correction (DDFM=KR), which is based on the Satterwaith method, should be routinely used. These corrections will usually (but not always) result in non-integer estimated degrees of freedom. Presence of only integer degrees of freedom in a given analysis is a likely indicator that the

degrees of freedom were not appropriately adjusted. Obviously, correctly estimated degrees of freedom are essential for PROC MIXED to provide correct *F*-tests of hypotheses.

### **5. Use appropriate adjustments in mean separations, and slices.**

The LSMEANS statement provides a method of examining differences among means corresponding to levels of a treatment. The default *P*-values corresponding to each of these pairwise comparisons are not adjusted to control the experimentwise error rate. Therefore, the overall chance of a Type I error (the error rate associated with declaring a significant difference that is actually due to random chance) increases with each additional comparison. Use of an appropriate correction (the ADJUST=TUKEY option of the LSMEANS statement, for example) will control the overall experimentwise error rate for these comparisons. When levels of interaction terms are examined, the numbers of means may become quite large, and many of the combinations examined may be of little or no interest. Use of the TUKEY option in this case results in considerable loss of power, while use of no correction results in severe inflation of the experimentwise Type-I error rate. In these cases, the SLICE option of the LSMEANS statement provides a means of simultaneously testing for differences among levels of one factor within levels of a second factor. Thus, use of this option minimizes the inflation of Type-I error rates while maintaining a relatively high level of power in the test.

Finally, authors seldom explicitly report the models used in their analyses. With the use of PROC MIXED becoming more frequent (and rightly so), it is imperative that authors indicate the terms listed in both the MODEL and RANDOM statements of the analyses. Without this information, it is generally impossible for an editor, reviewer, or reader to assess the appropriateness of a given analysis. PROC MIXED represents a substantial advancement in analytical methodology that should become standard for most entomological uses. However, the benefits of its adoption will be limited until users recognize the key concepts that are essential for its valid application.

### **Useful References**

1. Bergerud, W. 1996. Displaying factor relationships in experiments. *Am. Statist.* 50: 228-233.
2. Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. *SAS System for Mixed Models*. SAS Institute, Cary, NC.
3. Milliken, G. A., and D. Johnson. 1992. *The Analysis of Messy Data*. Vol. 1: *Designed Experiments*. Chapman & Hall, NY.
4. SAS Institute. 1999. *SAS OnlineDoc*, V8. SAS Institute, Cary, NC.