CellPress
REVIEWS

Opinion

Mixed Models Offer No Freedom from Degrees of Freedom

Göran Arnqvist[1,*]

Statistics matter greatly in biology, whether we like it or not. As a discipline with an empirical inclination, we are faced with data every day and we rely on inferential statistical models to make sense of it and to provide us with novel insights. Much of the time, the growing level of complexity and sophistication of the models we put to use in ecology and evolution have led to more appropriate analyses of our data. However, this is not always the case. Here, I draw attention to a classic flaw in inferential statistics that has resurfaced in a new flavor as a result of increased reliance on complex linear mixed models – the multifaceted and disturbingly persistent problem of pseudoreplication.

Highlights

Statistical errors due to pseudoreplication are still very common in primary studies.

The use of linear mixed models to analyze data can ideally resolve such problems, but only if models are structured correctly.

The use of linear mixed models can instead generate such problems, if models are not appropriately structured.

Inappropriate use of linear mixed models is very common in ecology and evolution, and we need to strive toward a more informed use of such models when analyzing our data.

The Problem of Pseudoreplication

When Hurlbert [1] alerted us to the problem of **pseudoreplication** (see Glossary) in ecology back in 1984, it was rampant and often obvious. He found that pseudoreplication occurred in about half of all published studies that used inferential statistics to analyze field experiments. While statistical errors due to pseudoreplication may possibly have decreased somewhat in frequency among studies of that particular sort [2], this appears not to be the case in other domains of biology where pseudoreplication is still very common and where its frequency is comparable to that documented by Hurlbert back in 1984 (e.g., [3–9]).

Pseudoreplication (Box 1) can be defined as testing for effects with an error term inappropriate to the hypothesis being considered [1] or, more generally, relating effects to inappropriate error variances. It reflects the wider issues of determining what constitutes the appropriate level or scale of independent replication for a given effect and occurs when we fail to account for dependencies across observations that are assumed by our models to be independent. It often arises as a result of a nested or hierarchical data structure and/or spatial or temporal dependencies of data. Modeling hierarchical data by complex **linear mixed models** (LMMs) can be a remedy for inferential errors due to pseudoreplication (e.g., [10–13]). However, importantly, this will only ever be true if appropriate models are fitted to data (e.g., [6,13–15]). If **random effects** structures are not modeled correctly, then LMMs offer no antidote. In fact, the now widespread use of such models seems to instead have aggravated the general problem of pseudoreplication in biology (Box 1).

The Issues at Hand

It may seem paradoxical that more sophisticated statistical modeling should be associated with more problematic analyses. This concern stems from at least five circumstances.

First, fitting models with a correct random effect structure is often very difficult [6,13–15] and the inner workings of complex LMMs are more or less obscure to many of us that put them to use [10]. For example, many of us believe that simply including one or more main random effect terms will miraculously account for dependencies in data and yield appropriate assessments of all **fixed effects** and interaction terms. This is typically not the case in LMMs [6,15] (Box 1).

[1]Animal Ecology, Department of Ecology and Genetics, Uppsala University, Norbyvägen 18D, SE75236 Uppsala, Sweden

*Correspondence:
Goran.Arnqvist@ebc.uu.se (G. Arnqvist).

Because our understanding is often incomplete, the use of complex LMMs can bestow a deceptive air of authority upon unsound analyses.

Second, misspecification of the random effect structure leads to pseudoreplication and inappropriate deflation of *P*-values (and higher **type I error** rates). This overconfidence can stem from inflated test statistics, for example, by *F*-ratios that are formed by incorrect denominator mean squares, and also from improper **degrees of freedom** when test statistics are converted to *P*-values. It is important to recognize that LMMs offer no magic freedom from degrees of freedom and generally no mysterious dramatic elevation of statistical power [16].

Box 1. Pseudoreplication

Pseudoreplication occurs when we analyze our data as if we had more degrees of freedom than we actually do. Here, degrees of freedom can be thought of as the number of independent pieces of information available to us when estimating another piece of information. The spirit of most statistical tests essentially involves relating an observed effect that we are interested in to some appropriate measure of intrinsic, natural random variation and asking how likely it is that our observation is the result of chance. This is most obvious in *F*-tests, where *F* represents a ratio of explained variance to unexplained variance or, alternatively, between-group variability to within-group variability.

Let us say that we have measured some response variable, say reproductive output (denoted *Y* in Figure I), in 20 individuals in each of 10 different populations (total $N = 200$ individuals). Some populations are blue and some are orange, representing, for example, an experimental treatment or an environmental dichotomy (color). To assess whether the 10 populations differ from one another (H_0: $\mu_1 = \mu_2 = \ldots = \mu_{10}$, where μ is the mean for that population), differences in population means are evaluated over the within-population variance (σ^2_w). This variance equals the error or sampling variance, which is best estimated by all 200 individuals measured. Hence, the error degrees of freedom will therefore reflect the total number of individuals measured in all populations. The logic is intuitive: if we ask if populations differ at all, then differences between populations should be related to variation within populations and all individuals measured provide information on this important metric. Individuals are the appropriate level of independent replication.

If we instead ask whether there is an effect of color (H_0: $\mu_{1–5} = \mu_{6–10}$), then the difference between color groups should be related to the among-population within-group variance (σ^2_a). This variance then equals the error variance. Hence, the error degrees of freedom should reflect the number of populations measured. Populations are the appropriate level of replication. The logic is again intuitive: if we ask whether blue and orange populations differ, then any differences between these groups should be related to variation among populations within color. Relating it to variation across individuals within populations represents a form of pseudoreplication.

These simple principles apply equally when evaluating data with complex LMMs, no matter how these models are fitted to data and which inferential strategy is employed. In such models, color would be modeled as a fixed effect factor along with other fixed effects and population identity (i.e., subject) is modeled as a random effects factor. Any test of effects of color should be evaluated over variation among populations, not over variation among individuals.

Importantly, this also applies to interaction terms containing color. If we, for instance, ask whether the relationship between *Y* and some factor or continuous variable *X* (say the relationship between reproductive output and body size) is different in populations of different color (i.e., a $X \times$ color interaction), the number of independent observations of this relationship (top two panels of plots in Figure I) equals the number of populations (i.e., 10) and not the number of individuals (i.e., 200). The analysis and the error degrees of freedom should reflect this [6].

If the effect of color (whether main effects or interactions) would be evaluated over within- rather than among-population variance, this can lead to deflated standards errors and confidence/credible intervals, inflated test statistics, inflated degrees of freedom for the error term, and consequently, to deflated *P*-values for fixed effects [6].

Unfortunately, pseudoreplication stemming from inappropriate error terms in mixed models is very common in ecology and evolution [3–9]. For example, an inspection of all papers utilizing 'experimental evolution' and published in the *Journal of Evolutionary Biology* during 2015–2019 ($N = 40$) revealed 11 cases of appropriate modeling, eight cases where the analyses were at least partly inappropriate, and 21 cases where inferential models and methods were not given in sufficient detail to assess whether the analyses were appropriate (many of which were likely inappropriate).

Note that in the simple example above, 'populations' (i.e., subjects) does not necessarily have to represent 10 biological populations made up by many individuals but could denote any hierarchical random effect factor, for example, 10 randomly selected individuals for which many measures of *Y* have been recorded.

Glossary

Cell mean: a cell in a design represents a unique combination of levels of all fixed and random factors, for example, assaying condition number three in population number seven in the example in Box 2 (in which there are 30 cells). The cell mean is simply the mean of all replicate observations in that cell.
Degrees of freedom: the number of values in the calculation of a metric that are free to vary.
Fixed effect: a model parameter for which the levels or values are assumed to be fixed (i.e., constant). The actual levels of a fixed effect factor are typically set or determined by the experimenter and fixed effects are estimated as, for example, differences between group means or as a slope.
Linear mixed model: a statistical model that is linear in its structure, wherein a focal response variable is considered a function of several variables and/or factors some of which have fixed and others random effects.
Pseudoreplication: occurs when analyzing data as if one had more independent samples, observations, or replicates than is actually the case.
Random effect: a model parameter assumed to represent a random variable. The actual levels of a random effect factor are considered a random subset of an infinite number of possible levels and random effects are estimated as variances or covariances.

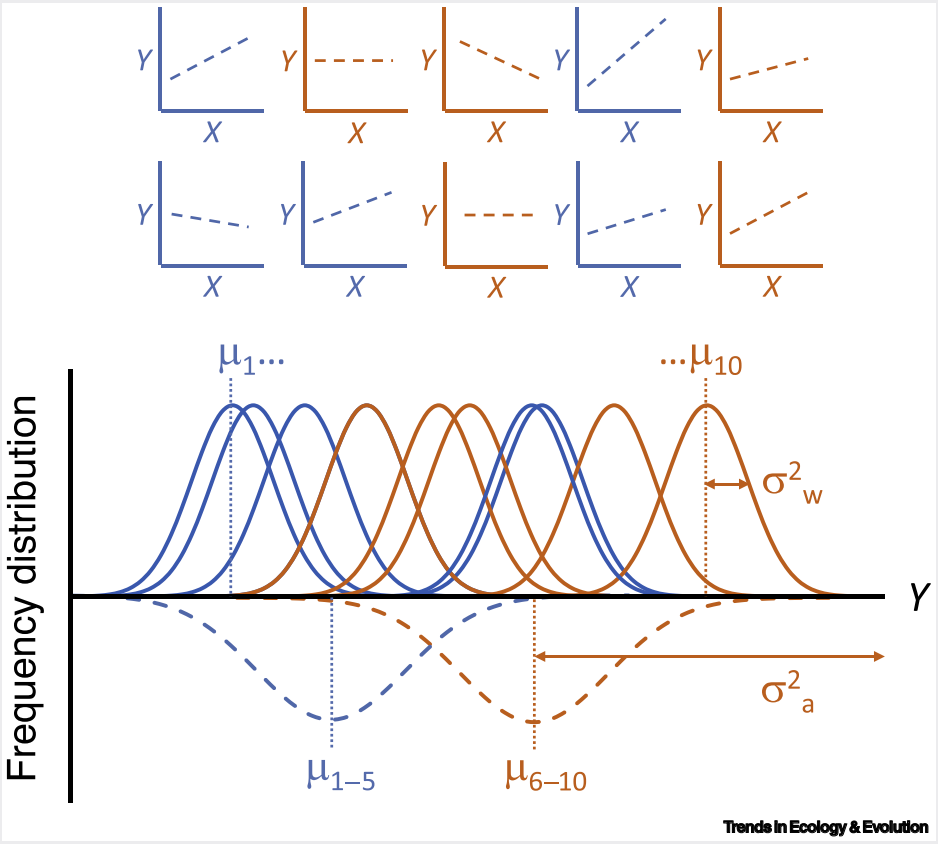

Figure I. A Schematic Illustration of Data from 10 Different Populations. Populations belong to either a blue or an orange category, that differ in the mean value (μ) of some response variable *Y* that shows a common within-population variance (σ^2_w). Each category of five populations also shows a grand mean value (e.g., $\mu_{1–5}$) that shows a common among-population variance (σ^2_a). The 10 populations also differ in the manner in which a covariate (*X*) relates to the response variable, as illustrated at the top.

Third, due to the complexity of data and of the mixed models fitted, analytical methods and strategies are often not transparent and are often not given in sufficient detail. This can make it impossible for reviewers to assess whether appropriate inferential models were used in primary research contributions and, thus, to help rectify problematic analyses [5,6].

Fourth, assessing the potential gravity of pseudoreplication in LMMs has been made even more difficult, or even impossible, as alternative strategies to model fitting and evaluation have gained ground. Traditional tests of focal fixed effects in mixed models are done by *F*-tests, which are transparent: they allow the direct assessment of whether an analysis may suffer from pseudoreplication as both the effect (numerator) and the error (denominator) degrees of freedom accompanies the *F*-ratio (Box 1). However, because data can be unbalanced and/or error distributions not normal, we often and rightly turn to alternative model fitting strategies which rely on test statistics (e.g., log-likelihood ratios) that lack transparency: they offer no information on the level at which data are regarded as being independently replicated. Again, this can make it impossible to assess how appropriate a given analysis is.

Fifth, pseudoreplication is sometimes believed to be a problem that emerges only when we engage in explicit null hypothesis testing, and that inferences based on parameter estimates and their confidence intervals (or the Bayesian equivalent, credible intervals) are somehow immune to these issues. They are not. Here, a misspecification of the random effects structure can generate inferential overconfidence by inappropriate shrinkage of the inferential intervals. The models employed in such analyses are often complex and opaque, especially in Bayesian model fitting strategies. It can be difficult or even impossible to work out at what level of replication inferential intervals were estimated. Yet, an appropriate random effect structure is as essential here as it is when engaging in more traditional model fitting and hypothesis testing strategies [15].

Some Simple and Concrete Advice

1. Explicitly state your inferential model in the paper or, at the very least, in supplementary information. This is often best done in tables, presenting estimates of all estimated parameters in the model (random and fixed) accompanied by the tests or confidence intervals that you rely upon, but can also include scripts detailing fitted models. This is required for transparency in the primary literature [17].
2. Whenever possible, present or at the very least inspect approximate *F*-ratios and their associated numerator and denominator degrees of freedom even when you lean on other test metrics (e.g., χ^2 or confidence intervals) for your final and main inferences. When using generalized LMMs [18], compare the results with analogous conventional LMMs that can be evaluated by *F*-tests. This can serve as a sanity check: it allows you to assess whether the error degrees

Box 2. An Example

We can use a relatively simple hypothetical data set to illustrate that data can be modeled in a number of different ways and that choice of inferential model can matter. We have collected data on *Y* from 10 different populations (POP), of which five belong to one treatment/sort and five to another. From each population, we also have data of three different types (e.g., assaying conditions) and we have data for 20 individuals of each type in each population. We thus have in total 10 × 3 × 20 = 600 observations of *Y*. However, these 600 observations are clearly not independent and replication here should reflect the number of populations (i.e., subjects) involved. We are interested in the fixed effects of our treatment (T), of the effects of assaying condition (AC), and of whether treatment effects might differ across assaying condition (T × AC; Table I).

A: All terms are here modeled as fixed effects, as if all 600 observations were independent.
B: An LMM including the random effect of POP, but where POP has not been appropriately nested within T.
C.1: An LMM including the random effect of POP, where POP has been nested within T. This is equivalent to a 'random intercepts' model. Here, mean *Y* is assumed to only vary in magnitude across populations within T, in the same concerted manner for all three levels of AC. Note that the effects of AC and T × AC are here evaluated over individuals rather than over populations, which is evident from the denominator degrees of freedom.
C.2: The same model structure as in C.1, but instead evaluated using Wald χ^2 tests.
D.1: An LMM instead including two random effect components: POP within T and the interaction between AC and POP within T. This essentially allows the overall effect of AC to vary across populations. This is in spirit a simple 'random slopes' model.
D.2: A model with the same structure as in D.1, but instead evaluated using Wald χ^2 tests.
E: A repeated measures ANOVA of cell means (10 × 3 = 30 observations), regarding POP as random effects subjects. These types of models yield appropriate tests of fixed effects, but cannot be used when some cells in the design are empty or when the within-subjects variable of interest is a covariate (rather than a factorial variable, as AC is here).
F: An LMM including six explicit random effect components; each of the three AC levels separately as well as the covariance between them. This represents what could be called a full 'random slopes' model, as it allows the shape of the effect of AC to vary across populations within T.

As is illustrated by this simple hypothetical example, what models we fit to data and how we choose to evaluate them can affect the conclusion we draw [6,9,15]. In the current example, models D.1, E, and F can all be considered appropriate. Models of the type represented by C.1 and C.2 are inappropriate but are common in ecology and evolution.

Note that if each observation is associated with a value of a covariate *X*, rather than a level of a factor as in this example (i.e., AC), then additional efforts may be needed to separate the effects of the covariate within each subject from those across subjects [14]. Data and examples scripts are available as Supplementary Information.

Table I. Different Models Fitted to the Example Data Set.

Model	Effect	Numerator degrees of freedom (ndf)	Denominator degrees of freedom (ddf)	*F*	*P*	Wald χ^2	*P*
A	T	1	594	5.27	0.022		
	AC	2	594	2.57	0.077		
	T × AC	2	594	3.40	0.034		
B	T	1	586	5.05	0.025		
	AC	2	586	2.57	0.077		
	T × AC	2	586	3.40	0.034		
C.1	T	1	8	5.05	0.055		
	AC	2	586	2.57	0.077		
	T × AC	2	586	3.40	0.034		
C.2	T	1				5.05	0.025
	AC	2				5.15	0.076
	T × AC	2				6.80	0.033
D.1	T	1	8	5.05	0.055		
	AC	2	16	1.64	0.226		
	T × AC	2	16	2.16	0.148		
D.2	T	1				5.05	0.025
	AC	2				3.27	0.195
	T × AC	2				4.32	0.115
E	Between subjects						
	T	1	8	5.05	0.055		
	Within subjects						
	AC	2	16	1.64	0.226		
	T × AC	2	16	2.16	0.148		
F	T	1	8	2.77	0.135		
	AC	2	7	1.76	0.240		
	T × AC	2	7	1.53	0.280		

of freedom (reflecting replication) and the estimated effects in your inferential model are reasonable for all terms. For example, some χ^2-based tests such as Wald tests can sometimes provide unreasonable evaluations of LMMs (Box 2).

3. As another sobering exercise, evaluate your data also using much simpler models of group means. This could involve anything from simple *t*-tests to partly nested mixed models of **cell means** [e.g., repeated measures analysis of variances (ANOVAs)]. If you find much larger or more significant fixed effects in your complex LMM than in the simpler tests, you have likely not modeled the random effects structure in an appropriate manner [6]. Be wary.
4. When you are interested in modeling interactions between fixed effects that apply within and between your random subjects, then include random parameters that allow the within-subjects effect to vary in shape across subjects (i.e., fit random slopes models). This is typically done by estimating several variance and covariance components. This allows intercepts and 'slopes' to differ, and even to covary, across levels of your random factor (e.g., populations) and it results in appropriate confidence intervals, error variances, and degrees of freedom for fixed effects [15]. Failure to allow within-subjects effects to vary across subjects is a particularly common problem in our domain [6].

5. Do not perform model simplification or reduction of random terms prior to arriving at your inferential model, where nonsignificant effects (including random interactions) that reflect levels of replication or design in your data are removed. This is because dropping random terms can greatly affect tests of fixed effects in a manner resulting in pseudoreplication [6]. Trust the full model – the random effects structure should be determined by your design rather than by your data [15]. Note that random effects (i.e., variances or covariances) that are inestimable or are negative are sometimes dropped automatically when fitting models, which can result in pseudoreplication.
6. Use restricted maximum likelihood (REML) estimation rather than maximum likelihood to fit your model, as REML yields unbiased estimates under a wider range of conditions. This is important if your data are imbalanced.
7. If your design is factorial, consider instead analyzing cell means using a repeated measures ANOVA (i.e., a partly nested mixed model). This is only possible if there are no empty cells in your data, but it is more likely to yield analyses with appropriate structure and degrees of freedom.
8. Strive to have a reasonable number of levels (at the very least, say, four to five subjects) of your random effects within each group (i.e., five populations in Box 1). Although there is some controversy surrounding this issue and how it affects estimates of fixed effects, random effects are modeled as a variance in mixed modeling and estimating a variance parameter accurately based on very few observations (i.e., subjects) is certainly problematic [19,20]. If you estimate several random effects in your model, this potential problem is aggravated.

Concluding Remarks

It is clear that overconfidence and type I statistical errors are common in LMMs of data in our domain, as a result of pseudoreplication. Complex inferential models and model fitting strategies are sometimes very useful, as they potentially allow us to model our data more appropriately when data are problematic [13,17]. However, many of us frequently err. We therefore also need to fully acknowledge that more complex models place much higher demands on our statistical proficiency and we should take precautionary steps to ensure a more informed use of LMMs.

Acknowledgments

I am grateful to D. Berger and O. Leimar for discussions and comments on a previous draft of the manuscript. This work was funded by the European Research Council (GENCON AdG-294333), the Swedish Research Council VR (621-2014-4523) and FORMAS (2018-00705).

Supplemental Information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tree.2019.12.004.

References

1. Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecol. Monograph* 54, 187–211
2. Heffner, R.A. *et al.* (1996) Pseudoreplication revisited. *Ecology* 77, 2558–2562
3. Ramage, B.S. *et al.* (2013) Pseudoreplication in tropical forests and the resulting effects on biodiversity conservation. *Conserv. Biol.* 27, 364–372
4. Waller, B.M. *et al.* (2013) Pseudoreplication: a widespread problem in primate communication research. *Anim. Behav.* 86, 483–488
5. Lazic, S.E. *et al.* (2018) What exactly is 'N' in cell culture and animal experiments? *PLoS Biol.* 16, e2005282
6. Schielzeth, H. and Forstmeier, W. (2008) Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* 20, 416–420
7. Ramírez, C.C. *et al.* (2000) Pseudoreplication and its frequency in olfactometric laboratory studies. *J. Chem. Ecol.* 26, 1423–1431
8. Kroodsma, D.E. *et al.* (2001) Pseudoreplication in playback experiments, revisited a decade later. *Anim. Behav.* 61, 1029–1033
9. Lazic, S.E. (2010) The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* 11, 5
10. Millar, R.B. and Anderson, M.J. (2004) Remedies for pseudoreplication. *Fish. Res.* 70, 397–407
11. Paterson, S. and Lello, J. (2003) Mixed models: getting the best use of parasitological data. *Trends Parasitol.* 19, 370–375
12. Colegrave, N. and Ruxton, G.D. (2018) Using biological insight and pragmatism when thinking about pseudoreplication. *Trends Ecol. Evol.* 33, 28–35
13. Harrison, X.A. *et al.* (2018) A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* 6, e4794
14. Van de Pol, M. and Wright, J. (2009) A simple method for distinguishing within-versus between-subject effects using mixed models. *Anim. Behav.* 77, 753–758

15. Barr, D.J. *et al.* (2013) Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278
16. Murtaugh, P.A. (2007) Simplicity and complexity in ecological data analysis. *Ecology* 88, 56–62
17. Parker, T.H. *et al.* (2016) Transparency in ecology and evolution: real problems, real solutions. *Trends Ecol. Evol.* 31, 711–719
18. Bolker, B.M. *et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24, 127–135
19. Gelman, A. and Hill, J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press
20. Harrison, X.A. (2015) A comparison of observation-level random effect and beta-binomial models for modelling overdispersion in binomial data in ecology and evolution. *PeerJ* 3, e1114