Paper 349-2011

Living with Generalized Linear Mixed Models

Walter W. Stroup, Department of Statistics, University of Nebraska, Lincoln, NE, USA

ABSTRACT

In the 1980s, before PROC MIXED or PROC GENMOD, "linear models" meant the "general" linear model as implemented by PROC GLM. Three decades later, the meaning of "linear models" has fundamentally changed. The introduction of PROC GLIMMIX in 2005 was a watershed moment. Now "linear model" means "generalized linear mixed model." The notion that the "general" linear model was once considered "general" seems quaint. In living with PROC GLIMMIX over the past five-plus years, several issues have become apparent that were not issues during PROC GLM's heyday. One set of issues concerns inference space and conditional versus marginal modeling. When you learn about the GLMMs, you often find that you must first "unlearn" the old ways of thinking about models. The second set of issues concerns power, sample size, and planning. These areas have not caught up with GLMMs. Finally, applied statistics curriculum seems caught in a time warp, leaving students and especially consumers of statistical methods unprepared for contemporary statistical practice.

INTRODUCTION

Once upon a time "linear model" meant $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$. In certain technocratic circles, TLA means "three letter acronym." You know a procedure rates if it has a TLA. In the 1970s, the TLA for $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$ was GLM – "General" Linear Model. PROC GLM took its name from this TLA.

Once upon a time is no more.

Fast forward to 2011. In literate linear model circles, "GLM" is now the TLA for general*ized* linear models. What used the be the "General" Linear Model is now just the LM – by modern standards $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$ is not at all "general." It has been demoted and does not even warrant a TLA. Confusingly for SAS® users, PROC GLM cannot compute GLMs. For this, you need PROC GENMOD or PROC GLIMMIX. In 2011, the term "linear model" connotes General*ized* Linear *Mixed* Model (GLMM – sufficiently important to rate a *four* letter acronym!). All linear models, linear mixed models (PROC MIXED), modern GLMs (PROC GENMOD) and modern LMs (formerly GLMs) are special cases of the GLMM. To fully specify a GLMM we need

- a linear predictor: $\boldsymbol{\eta} = \mathbf{X\beta} + \mathbf{Zb}$
- two distributions: $\mathbf{y} \mid \mathbf{b} \sim (\boldsymbol{\mu} \mid \mathbf{b}, \mathbf{R})$ and $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$
- a link function: $g(\boldsymbol{\mu} \mid \mathbf{b})$; or alternatively, an inverse link function: $h(\mathbf{X\beta} + \mathbf{Zb}) = \boldsymbol{\mu} \mid \mathbf{b}$

For SAS users, the watershed moment came in 2005. On March 17, Oliver Schabenberger sent me an e-mail. Subject: "Rushing this message to you on my way out of town." Message begins, "The production release for PROC GLIMMIX is now available for download..." The GLMM had been an active area of research for over two decades. Macros had been available. Now, however, we had a full-fledged PROC with performance and features that clearly made it the undisputed flagship of linear model software.

Five years on, it has become clear that the GLMM represents more than just a set of advanced techniques for high-end practitioners only. What we have here is something far more fundamental. For the first time in several decades – probably since the convergence of matrix algebra, third generation computers and linear models made modern statistical computing possible – we have a full-blown challenge to established paradigms regarding what defines "standard statistical practice" and what constitutes "the basics" in the core curriculum of statistics.

When PROC MIXED came out, we could essentially do the same old thing, only better, especially with split-plots and repeated measures. We just added $\mathbf{Zb}$ to $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$, added flexibility in specifying $\mathbf{R} = Var(\mathbf{e})$, and carried on, our mindset intact. The GLMM requires a more comprehensive change. In this paper, we consider three aspects of this challenge.

The first concerns the way we approach bread-and-butter design and analysis, e.g. paired comparisons and randomized complete block designs, especially with non-Gaussian data. Nothing exotic: proportions and counts. Along the way we gain a new appreciation for what may have triggered R.A. Fisher's generally hostile reaction to early work in statistical modeling.

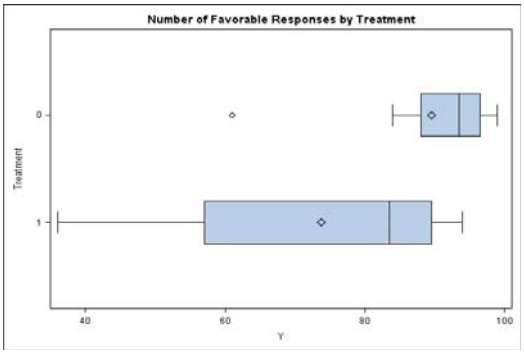
The second concerns the way we assess power and sample size when planning future studies. Existing methods generally reflect, for want of a better term "pre-GLMM thinking." Applied to commonplace situations that clearly call for GLMM thinking, conventional power and sample size methods can produce inappropriate – in some cases catastrophically inappropriate – assessment of design requirements.

Finally, the third challenge concerns our approach to core service classes in statistical methods course and our concept of core curriculum in graduate programs in statistics. As we go though the design-and-analysis and planning examples, we shall see that there are GLMM-driven ideas that never occurred to us to include in traditional courses but must now be considered "basic." The question is not *whether* we need to teach them in introductory courses but *how* are we going to teach them in digestible form in our introductory classes. "GLIMMIX is too hard" won't do.

THE FIRST CHALLENGE: BASIC ANALYSIS – OLD DOGS AND NEW TRICKS

Example 1: a paired comparison. We have 8 sites and 2 treatments. For each treatment at each site, we observe 100 subjects. Each subject either has a favorable or an unfavorable response. Let Y_{ij} denote the number of favorably responding subjects on treatment *i* at site *j* . Figure 1 shows box plots of the data for each treatment at the 8 sites.

Figure 1. Box Plot of Binomial Paired Comparison Data



Inspecting Figure 1, it seems obvious that the two treatments differ. How is analysis of these data taught in an introductory methods class? First, with 100 subjects per site-treatment combination, it *seems* that we could invoke the Central Limit Theorem: the sample proportion $p_{ij} = Y_{ij}/100$ should be approximately distributed $N\left(\pi_{ij}, \pi_{ij}\left(1-\pi_{ij}\right)/100\right)$. Therefore, we *should* be able to use the linear model $p_{ij} = \mu + \tau_i + \rho_j + e_{ij}$ where τ_i denotes the i[th] treatment effect and ρ_j denotes the j[th] pair(site) effect. If the observed sites represent a sample of a target population, we could assume the site effects are i.i.d. $N\left(0, \sigma_P^2\right)$ and use PROC MIXED or PROC GLIMMIX. Because this paper concerns GLIMMIX, the statements are

```
proc glimmix data=intro_binomial;
 class Site Treatment;
 Pct=Y/N;
 model Pct =Treatment;
 random intercept / subject=Site;
 lsmeans Treatment / diff;
```

Output 1. Binomial: Normal Approximation

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Treatment	1	7	3.28	0.1132

Treatment Least Squares Means		
Treatment	**Estimate**	**Standard Error**
0	0.8950	0.06379
1	0.7375	0.06379

<Living with Generalized Linear Mixed Models>, continued

Output 1 shows the relevant GLIMMIX listing. We know that analysis of variance yields the same results. We see two disturbing features. First, the *p*-value for $H_o: \tau_0 = \tau_1$ is 0.1132, which precludes concluding a statistically significant treatment effect on the likelihood of a favorable response. Second, the standard errors are equal despite the binomial distribution's well-known mean-variance relationship $\pi(1-\pi)$.

In pre-GLMM world, we would address the variance issue by transforming the data using the arc-sine-square-root. The GLIMMIX statements and output are not shown here; the estimates of the treatment 0 and treatment 1 means, back-transformed to the data scale, are 0.916 ± 0.040 and 0.760 ± 0.067 respectively and *p*=0.0605.

Now let's enter the modern world and use a GLMM. First, we specify the model:

- Linear predictor: $\eta_{ij} = \eta + \tau_i + \rho_j + (\rho\tau)_{ij}$
- Distributions: $\left[y_{ij} \mid \rho_j, (\rho\tau)_{ij}\right] \sim Binomial\left(100, \pi_{ij}\right)$; ρ_j i.i.d. $N\left(0, \sigma_S^2\right)$; $(\rho\tau)_j$ i.i.d. $N\left(0, \sigma_{ST}^2\right)$
- Link function: logit = $\log\left(\pi_{ij} / \left(1-\pi_{ij}\right)\right)$

Notice that a site × treatment interaction effect, $(\rho\tau)_{ij}$, appears in the linear predictor here, whereas it did not previously. More about that later. Also, notice what we do *not* do: we do not form the model from $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$. The required GLIMMIX statements are

```
proc glimmix data=intro_binomial;
 class Site Treatment;
 model Y/N =Treatment;
 random intercept Treatment / subject=Site;
 lsmeans Treatment / diff ilink;
```

Output 2. Binomial: GLMM

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Treatment	1	7	6.75	0.0355

Treatment Least Squares Means				
Treatment	**Estimate**	**Standard Error**	**Mean**	**Standard Error Mean**
0	2.5502	0.4405	0.9276	0.02958
1	1.2698	0.4232	0.7807	0.07245

The *p*-value for the GLMM test of $H_o: \tau_0 = \tau_1$ is 0.0355. In the table of **Least Squares Means** ESTIMATE shows the estimated logit; MEAN shows the estimated probability ($\hat{\pi}_i$) for each treatment. They are 0.928 and 0.781 respectively, compared to 0.895 and 0.738 obtained with the normal approximation. This is no accident. For reasons explained immediately below, the GLMM estimates will *always* be closer to 1 when $\hat{\pi}_i > 0.5$ and closer to 0 when $\hat{\pi}_i < 0.5$. Moreover, the GLMM test is more powerful, without compromising type I error control, so the greater *F*-value and lower *p*-value are not merely happenstance.

MARGINAL AND CONDITIONAL MODELS: WHAT YOU SEE ISN'T WHAT YOU GET

Example 1 highlights an underappreciated, often unrecognized aspect of working with data from non-Gaussian distributions. The normal-approximation ANOVA and the GLMM are not merely different approaches to the same problem. They target different parameters and only the GLMM actually targets a binomial probability. Consider the conceptual premise of this example. We have observations for a given site-treatment assumed to vary according to a binomial distribution. We also have variability among sites. What does "variability among sites" mean? Not an idle question: many statistics graduate students – even good ones – struggle to give a coherent answer. For modeling purposes, we are saying that the binomial probability is randomly perturbed from site-to site. Since we model the logit, a.k.a. the log-odds, variance among sites means site-to-site variation in log-odds.

<Living with Generalized Linear Mixed Models>, continued

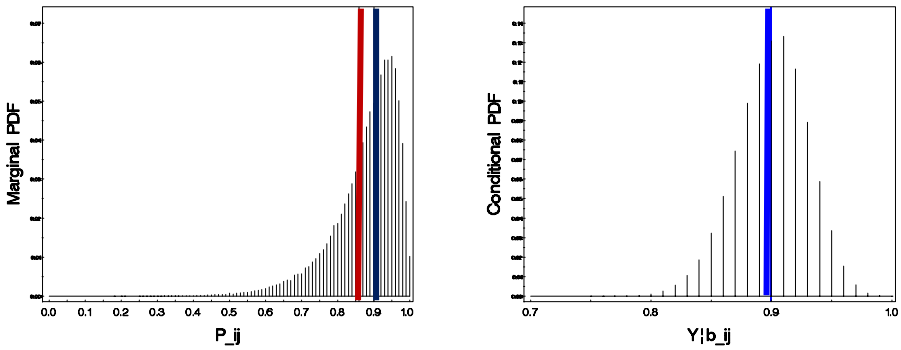
We never observe any of the distributions given in the GLMM model statement directly. The only direct observation we can make is on the response variable Y, whose distribution is given by the marginal p.d.f.

$f(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{b}) f(\mathbf{b}) d\mathbf{b}$, where $f(\mathbf{y}|\mathbf{b})$ is the binomial p.d.f. and $f(\mathbf{b})$ is the joint p.d.f. of the random site and site-by-treatment effects, ρ_j and $(\rho\tau)_{ij}$.

The marginal p.d.f. does not simplify much beyond what we have written here, but we can use simulation to visualize it. Figure 2 shows the empirical marginal $f(\mathbf{y})$ (left) and conditional $f(\mathbf{y}|\mathbf{b})$ (right) distributions from 10,000 simulated data sets assuming $f(\mathbf{y}|\mathbf{b})$ is binomial$(100, 0.9)$, $\sigma_S^2 = 0.5$ and $\sigma_{ST}^2 = 1$. The probability and variance components approximate our estimates from the GLMM analysis.

Figure 2. Distributions for Logit-Normal Model

a. Marginal **b. Conditional**

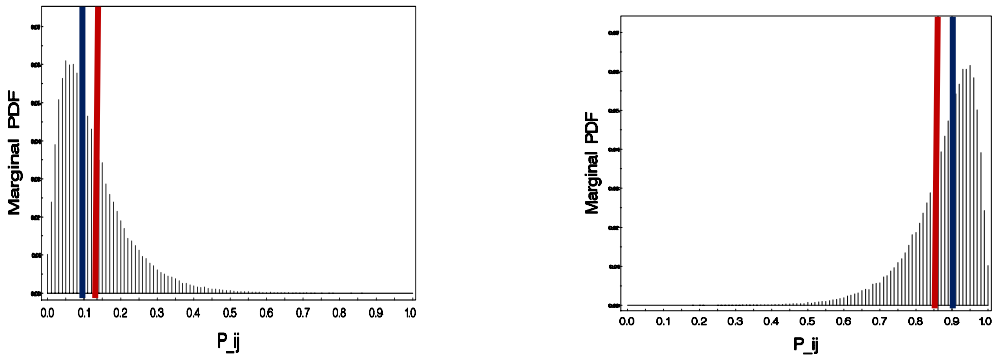


For the marginal distribution, the horizontal axis denotes the sample proportion. Notice that the blue bar corresponds to 0.90, the binomial probability $\pi = 0.9$, The red bar, at roughly 0.86, is the marginal expectation of the sample proportion. Compare Figure 2a with Figure 2b, the empirical p.d.f. of a true binomial$(100, 0.9)$. We see that the normal approximation applies to the (unobserved) conditional distribution, not the (actually observed) marginal distribution. With ANOVA, we use $\bar{p}$, the sample mean of the observed proportions, to obtain an unbiased estimate of $E(Y/N) = \mu_P$. With the GLMM, we estimate $\text{logit}(\pi) = \eta + \rho_i$ then use the inverse link to obtain $\hat{\pi} = 1/\left[1 + \exp(-\hat{\eta})\right]$. In the context of Figure 2, the normal approximation targets $\mu_P \cong 0.86$; the GLMM targets $\pi = 0.9$. Only the GLMM estimates a binomial probability.

Figure 2a shows that for any $\pi > 0.5$ the marginal mean is always less than the binomial probability because the marginal distribution is always left-skewed. For $\pi < 0.5$ the skewness reverses; the marginal mean will always be greater than the binomial probability. If we have a study comparing two probabilities, Figure 3 shows the nightmare scenario and its impact on our ability to accurately estimate odds-ratio and on power.

Figure 3. Conditional vs. Marginal Model Estimation of Treatment Difference

Treatment 0 **Treatment 1**



In Figure 3 the red vertical bars show marginal means and the blue vertical bars show the actual binomial probabilities for the two treatments. ANOVA targets the red bars; the GLMM targets the blue. These plots make it easy to see that the normal approximation, which seemed so reasonable when we introduced the example, actually lures us into an analysis that very likely is not what we intended.

UNLEARNING AND RELEARNING MODEL CONSTRUCTION

Speed (2010) published an *IMS Bulletin* commentary on the uneasy co-existence of ANOVA and linear modeling ever since Fisher's seminal work on ANOVA (Fisher and Mackenzie,1923). Noting Fisher's negative reaction to early work in statistical modeling ("confused" and "enraged" to quote Speed directly) Speed recommends a presentation by Yates (1935) to the Royal Statistical Society and in particular points to Fisher's comments following Yates' talk. Speed characterizes the split-plot as the "litmus test that separates those who understand ANOVA from those who don't get it." In a talk at the Joint Statistical Meetings, Oliver Schabenberger (2008), paraphrasing Vince Lombardi, said, "The split-plot isn't everything; it's the only thing."

Reading Fisher's comments produced an "ah ha!" moment: it occurred to me that following Fisher's analysis of variance *thought process* led to a coherent way to construct a GLMM and at the same time shed light on why the mindset implicit in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ so often paints us into an undesirable corner, as in Example 1. For my modeling and design classes, I've come up with a process I call "What would Fisher do?" or WWFD. More on that later.

In his comments, Fisher said every design has a "topographical" aspect and a "treatment" aspect. Fisher was discussing agricultural experiments; we can understand "topographical" broadly as the non-treatment components of a design structure – what design of experiments textbooks often refer to as the "experiment design" although the concept extends more generally to surveys and observational studies as well. Consider the design shown in Figure 4. I call it "the litmus test." It is a tweaked version of a design presented in Milliken and Johnson (2008).

Figure 4. The Litmus Test Design

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9	Block 10
0	0	0	0	0	3	3	3	3	3
1	1	1	1	1	4	4	4	4	4
2	2	2	2	2	5	5	5	5	5

Figure 4 shows an incomplete block design with 10 blocks, 3 units per block and 6 treatments denoted 0, 1, ..., 5. Introductory linear models classes, at least at the University of Nebraska, still focus heavily on the "general" linear model. In old-GLM world, BLOCK is the "topographical" aspect and TREATMENT is exactly what it says. The linear model my colleagues invariably assign is $y_{ij} = \mu + \tau_i + \rho_j + e_{ij}$, essentially the same model used in Example 1 for the normal approximation. My colleagues also ask who let this design see the light of day? It is a disconnected design rife with estimability problems, confounding, etc. Here we have a teachable moment. This hints at what may have driven Fisher to his contempt for much of the modeling work of his day.

WHAT WOULD FISHER DO?

First, the "treatment aspect" consists of more that TREATMENT. Clearly, we have two SETS, one containing treatments 0, 1 and 2, the other containing treatments 3, 4, and 5. Fisher suggested writing two skeleton ANOVAs (sources of variation and degrees of freedom), one for the topographical design and one for the treatment design, then integrating them. At that point, he contended, it should be clear how to proceed. Further, if one followed this process with requisite attention to detail, the process should work for designs of arbitrary complexity. Table 1 shows how it works in this example.

Table 1 . Skeleton ANOVA for Litmus Test Design

Topographical		Treatment		Combined	
Source	d.f.	Source	d.f.	Source	d.f.
		set	1	set	1
block	9			block|set	9-1=8
		treatment(set)	4	treatment(set)	4
unit(block)	20	"parallels" (Fisher's term)	24	unit(block) a.k.a. trt × blk(set) a.k.a. residual	20-4=16
Total	**29**	**Total**	**29**	**Total**	**29**

The topographical aspect consists of the 10 blocks and 3 units within each block. The placement of the treatment and topographic entries is important. In setting up the design, sets are randomly assigned to blocks, hence block is the unit of replication with respect to set. Treatments within sets are randomly assigned to units, thus unit(block) is the unit of replication with respect to treatment. Notice that set and treatment(set) each take degrees of freedom away from their respective units of replication in the final, combined ANOVA. We must work through this process to see clearly which effects must be random and which effects are fixed when we write the model.

At this point, we know that the unit of observation is the unit(block). In general, the unit of observation is last term in the combined skeleton ANOVA. We also know that the linear predictor needs to account for all of the other sources of variation in the combined ANOVA. How does this work?

The observations on the unit(block) have some probability distribution, generically denoted as $f(\mathbf{y}|\mathbf{b})$. In GLMM theory, $f(\mathbf{y}|\mathbf{b})$ may be any member of the exponential family or any quasi-likelihood. If we have Gaussian data and the observations are independent and homoscedastic, then $y_{ijk}|b(\alpha)_{ik} \sim NI\left(\mu_{ijk},\sigma^2\right)$, where *NI* denotes "normal and independent" and $b(\alpha)_{ik}$ denotes the ik^{th} block within set effect. On the other hand, we could have binomial proportions or counts, or any other non-Gaussian distribution in the GLMM stable. If we have counts, they could be Poisson, that is $y_{ijk}|b(\alpha)_{ik} \sim \text{Poisson}\left(\lambda_{ijk}\right)$, they could be negative binomial $y_{ijk}|b(\alpha)_{ik} \sim \text{NegBin}\left(\lambda_{ijk},\phi\right)$ etc.

The conditional distribution affects how we understand the last term in the skeleton ANOVA. With Gaussian data, μ_{ijk} and σ^2 are completely separate entities. The mixed model estimates μ_{ijk} but provides nothing regarding σ^2. The last entry in the ANOVA must be treated as "residual" because we need that information to estimate σ^2. For Gaussian data, the last term in the combined ANOVA corresponds to ***e*** in $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$. This, however, is the ***only*** case for which ***e*** in a linear model has this – or any – meaning. If we have Poisson observations, the variance equals the mean – once we estimate the mixed model, we have the mean *and* the variance among the units of observation. We don't need the residual to estimate variance. This also leaves it free for other purposes. On the other hand, if we have Negative Binomial observations, the variance among units depends partly on the mean and partly on a scale parameter. To estimate the latter, we need the information provided by the unit(block) source of variation. But it is not "residual" or "error" in the sense that we understand it with Gaussian data.

The take home message is this. If we understand the ANOVA *thought process* Fisher articulated and don't tie it to a specific set of arithmetic that is an artifact of the Gaussian distribution and computing technology limitations of a bygone time, we have a process that bypasses the "general" linear model and leads directly to the GLMM. The model resulting from this process for the Litmus Test design is

- linear predictor: $\eta_{ijk} = \eta + \alpha_i + \tau(\alpha)_{ij} + b(\alpha)_{ik}$, where α and τ denote set and treatment effects, respectively. In some cases, this linear predictor may be incomplete, depending on the distribution of the observations and how we understand the last term in the skeleton ANOVA. More about that below.
- distributions:
    - Observations: $y_{ijk}|b(\alpha)_{ik}$ as shown above.
    - Random model effect, i.e. blocks: $b(\alpha)_{ik}$ i.i.d. $N\left(0,\sigma_B^2\right)$
- link: $g\left(\mu_{ijk}\right) = \eta_{ijk}$. The specific link depends on the conditional distribution of the observations and possibly other considerations relevant to a given problem.

We now consider our second example to see how we implement these ideas with GLIMMIX.

EXAMPLE WITH COUNT DATA

Example 2: Suppose we have count data from a study that used the Litmus Test design. By count data, we mean that the observed response is the number of occurrences – e.g. number of weeds, number of microorganisms, number of defects, etc. – a non-negative integer. Historically, the Poisson distribution has typically been the default distribution for counts. In many areas – ecology, for example – alternative distributions, notably the negative binomial, often provide better models. The main issue with the Poisson distribution and count data is overdispersion. The Poisson entails a rigid and restrictive mean-variance relationship. Overdispersion occurs when the observed variance exceeds what the variance should be under the assumed distribution. In many disciplines overdispersion in count data is the rule, not the exception, and the magnitude of overdispersion can be substantial.

<Living with Generalized Linear Mixed Models>, continued

In our example, let us start by assuming a Poisson distribution and using the linear predictor given immediately above. We use the following GLIMMIX statements

```
proc glimmix data=a method=laplace;
 class block set trt;
 model y=set trt(set) / d=poisson;
 random intercept / subject=block;
 lsmeans trt(set) / ilink;
```

We must use METHOD=LAPLACE or METHOD=QUAD in order to get the conditional fit statistics, the appropriate diagnostic statistics for overdispersion. Output 3 shows relevant output.

Output 3. Selected GLIMMIX Results – Poisson Model without Block x Trt Effect

Fit Statistics for Conditional Distribution	
Pearson Chi-Square / DF	2.37

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
set	1	16	0.06	0.8151
trt(set)	4	16	7.79	0.0011

trt(set) Least Squares Means					
set	trt	Estimate	Standard Error	Mean	Standard Error Mean
0	0	1.6350	0.5071	**5.1296**	**2.6015**
0	1	2.2522	0.4959	**9.5086**	**4.7157**
0	2	1.4768	0.5112	**4.3790**	**2.2387**
1	3	1.6242	0.5054	**5.0746**	**2.5645**
1	4	2.2448	0.4961	**9.4387**	**4.6828**
1	5	1.9889	0.4993	**7.3074**	**3.6485**

Under **Fit Statistics for Conditional Distribution,** Pearson $\chi^2/df = 2.37$. In theory, $\chi^2/df = 1$ indicates complete absence of overdispersion, i.e. the data fit the Poisson assumption that the among-unit variance equals the mean. Pearson $\chi^2/df >> 1$ constitutes evidence of overdispersion.

In GLM literature, the suggested fix for overdispersion is to add an overdispersion scale parameter. You do this in GLIMMIX by adding a `random _residual_` statement. There are several problems with this. First, doing so creates a quasi-likelihood – you have a likelihood with the same form as the Poisson likelihood, except it has an additional parameter so that $Var\left(y_{ijk} \mid b(\alpha)_{ik}\right) = \phi\lambda_{ijk}$ instead of λ_{ijk}. This corresponds to no actual probability distribution *and* this approach implicitly targets the marginal distribution, creating the problems similar to those we saw in the first example. The second problem is that several studies, e.g. Young, et al (1998), have documented the ineffectiveness of the scale parameter approach in controlling type I error rate.

This takes us back to “what would Fisher do?” Unlike the normal distribution, the Poisson is a one-parameter distribution. This, combined with the linear predictor we’ve used, means we are implicitly assuming that all units within blocks are identical. That is, we assume that λ_{ijk} may be affected by block main effects and treatment, but individual units are otherwise identical. In reality, units probably do have unique characteristics and, rather than staying absolutely constant, it is more likely that λ_{ijk} is randomly perturbed from unit to unit within a block. We saw this with the site × treatment effect in Example 1. In a Gaussian model, we account for this random perturbation via σ^2. A Poisson has no analog of σ^2 - our GLMM model essentially leaves variability coming from the last line in the skeleton ANOVA unrepresented. The quasi-likelihood scale parameter is inadequate partly because it is on a linear scale (ϕ vs. the Gaussian σ^2) but mostly because it ignores design architecture the skeleton ANOVA makes explicit.

<Living with Generalized Linear Mixed Models>, continued

What happens when we follow the skeleton ANOVA? We amend the linear predictor by adding a term to represent unit(block) , yielding $\eta_{ijk} = \eta + \alpha_i + \tau(\alpha)_{ij} + b(\alpha)_{ik} + \tau b(\alpha)_{ijk}$ where $\tau b(\alpha)_{ijk}$ i.i.d. $N(0, \sigma^2_{BT})$. The GLIMMIX statements for this model are identical to the previous program, except for the revised RANDOM statement

```
random intercept trt(set) / subject=block;
```

Output 4 shows selected results.

Output 4. Selected GLIMMIX results – Poisson model with BLOCK x TREATMENT(SET)

Fit Statistics for Conditional Distribution	
Pearson Chi-Square / DF	0.22

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
set	1	16	0.08	0.7808
trt(set)	4	16	0.89	0.4950

trt(set) Least Squares Means					
set	trt	Estimate	Standard Error	Mean	Standard Error Mean
0	0	1.2650	0.5873	**3.5430**	**2.0809**
0	1	1.9914	0.5576	**7.3256**	**4.0844**
0	2	1.7521	0.5558	**5.7669**	**3.2054**
1	3	1.5270	0.5743	**4.6042**	**2.6443**
1	4	1.9823	0.5589	**7.2597**	**4.0578**
1	5	2.0838	0.5529	**8.0352**	**4.4424**

Notice the impact on the test statistics – in particular, for TREATMENT(SET). The *F*-value was 7.79; now it is 0.89. This is the characteristic impact of overdispersion. If not accounted for, it can severely inflate type I error rates. In this case, the estimates of the Poisson rate, given under MEAN in the **Least Squares Means** listing, do not vary appreciably among treatments, but the model not accounting for BLOCK x TREATMENT (SET) yields a wildly inflated *F*-value. Here, BLOCK x TREATMENT (SET) acts as the overdispersion parameter, and it does so via a legitimate probability distribution and in context of the design that produced the data.

We can formally evaluate overdispersion using a COVTEST statement to test $H_0 : \sigma^2_{BT} = 0$. The GLIMMIX statement

```
covtest 'blk x trt(set)' . 0;
```

produces Output 5.

Output 5. COVTEST Listing for Test of Overdispersion Variance Component

Tests of Covariance Parameters Based on the Likelihood					
Label	DF	-2 Log Like	ChiSq	Pr > ChiSq	Note
blk x trt(set)	1	227.64	33.54	<.0001	MI

The likelihood ratio $\chi^2 = 33.54$ with a *p*-value <0.0001, strong evidence supporting our decision to include BLOCK x TREATMENT(SET).

An alternative way to model the contribution of unit(block) is to change the assumed distribution $f(\mathbf{y} \mid \mathbf{b})$ from Poisson to a counting distribution with a scale parameter, such as the negative binomial or the generalized Poisson. The generalized Poisson requires adding user supplied statements to define the log-likelihood. Example 38.14 in the

GLIMMIX online documentation for SAS®/STAT shows the required statements. The negative binomial only requires you to change the DISTRIBUTION= (D =) option in the MODEL statement. The revised statements are

```
proc glimmix data=a method=laplace;
 class block set trt;
 model y=set trt(set) / d=negbin;
 random intercept / subject=block;
 lsmeans trt(set) / ilink;
```

Notice that we leave the BLOCK x TREATMENT(SET) effect out of the RANDOM statement, returning instead to the original linear predictor. It would be redundant to estimate the negative binomial scale parameter *and* σ^2_{BT}.

This concludes the first part of this paper. The take-home message here is that as the GLMM becomes the mainstream linear model, we need a paradigm shift in the way we conceptualize models. The $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$ mindset paints us into corners with no easy escape. Instead, we need to return to ANOVA's roots – the *thought process* not the ossified arithmetic part – and follow that process in view of contemporary theory and computing capability. This is the first challenge living with GLIMMIX has made obvious.

THE SECOND CHALLENGE: POWER, SAMPLE SIZE AND PLANNING IN A GLMM WORLD

Littell, et al (2006) devote a chapter to the use of PROC MIXED to compute power for linear mixed models. The method has its roots in work by Littell (1980), O'Brien & Lohr (1984) with PROC GLM. We can easily extend these methods to GLMMs using PROC GLIMMIX. The basic outline of the process:

1. Create an "exemplary data set" (Ralph O'Brien's term). This is a data set with the same structure we will eventually use to analyze the data once they are collected. In the exemplary data set, expected responses appear in place of observed data. These expected responses should reflect the minimum difference between treatments considered important. For example, in pharmaceutical testing, the minimum "clinically relevant" difference defines the expected response. Our objective is to identify a design capable of showing a "clinically relevant" difference to be statistically significant at a specified α-level with acceptably high power.
2. Run the intended analysis with GLIMMIX on the exemplary data using the model we intend to use for the observed data. In this step we must hold the covariance components constant. This requires knowledge of plausible values of these covariance components.
3. GLIMMIX computes approximate *F*-values. Multiplying them by the numerator degrees of freedom, **NumDF*FValue** in the SAS data set we create during the GLIMMIX run, yields the non-centrality parameter of the non-central *F*-distribution given the expected responses in the exemplary data set, the assumed covariance components and the design we are evaluating.
4. Use probability functions FINV and PROBF to determine the power associated with the design being evaluated. We use FINV to determine the critical value of the *F*-test given α and the numerator and denominator degrees of freedom. We then use PROBF to compute the probability that the test-statistic exceeds the critical value under the non-central *F* determined in (3).

Examples of all of the needed statements are given below as we work through a scenario with binomial response.

THE GLMM CHALLENGE TO POWER AND SAMPLE SIZE METHODLOLOGY

Simply put, current power and sample size software and methodology – i.e. the approach that's taught in statistical method courses for research workers – is woefully inadequate, even for mundane jobs like a 2-sample paired comparison with binomial (or count) data. In many cases, what we get with standard power and sample size methods is just plain wrong – sometimes catastrophically so. The problem is that conventional power and sample size methodology does not take into account the impact of random model effects and, for non-Gaussian mixed models, the combined impact of random model effects and non-normality. These impacts can be enormous.

If we are going to use power analysis effectively to plan research designs, we need to align the power we compute to the probability environment where the design actually lives. In an era of pervasive budget stress on research and development – in the private sector, in academia, you name it – this issue has never been more important.

Example 3. To see how this works, suppose we have a "standard" treatment, we'll call it **treatment 0**, known to yield a favorable response approximately 15% of the time. We have three experimental treatments, call them **treatments 1, 2 and 3**, that are claimed to increase the favorable response rate to at least 25%, perhaps even as high as 35%. We want to test this claim.

We have a binomial response. Our design problem amounts to determining how many subjects we need to

show $\pi_i - \pi_0 \geq 0.10$, or the equivalent odds-ratio or relative-risk, for a given α-level. We know that the variance of a binomial random variable increases as π approaches 0.5. We also know that our primary focus is on the difference between 0.15, π_0 for the standard treatment, and 0.25, the minimum π_i considered to be clinically relevant for any of the experimental treatments. We could either base our power analysis on the difference between $\pi = 0.15$ and $\pi = 0.25$ or we could hedge our bets and use the difference between $\pi = 0.25$ and $\pi = 0.35$ (because the higher variance will force a somewhat larger sample size).

We can determine the required sample size using PROC POWER. The needed statements are

```
proc power;
 twosamplefreq
   test=pchi
   proportiondiff=0.10
   /*relativerisk=1.4*/
   /*oddsratio=1.615385*/
   npergroup=.
   power=.8
   refproportion=.25;
```

Notice that we can express the "minimum clinically relevant" difference using either PROPORTIONDIFF, RELATIVERISK or ODDSRATIO. We get the same results. Here, we use PROPORTIONDIFF and comment out the alternatives. Output 6 shows the result.

Output 6. PROC POWER result for Binomial Treatment Comparison

Computed N Per Group	
Actual Power	**N Per Group**
0.801	329

According to Output 6 we need at least 329 subjects per treatment if $\alpha = 0.05$ and we want power to be at least 0.80. Now the research team throws us a curve. The researchers do not want to risk having their findings dismissed as location-specific, so they need this to be a multi-site trial. Also, they simply cannot handle more than 2 treatments or approximately 250 subjects at any given location. Some brainstorming produces three possible designs.

Figure 5. Three Possible Designs for 4-Treatment Binomial Comparison

a. Balanced Incomplete Block

Location	Treatments	
1	0	1
2	0	2
3	0	3
4	1	2
5	1	3
6	2	3

b. Control vs. Experimental

Location	Treatments	
1	0	1
2	0	1
3	0	2
4	0	2
5	0	3
6	0	3

c. Randomized Complete Block

Block	Location	Treatments	
1	1	0	1
	2	2	3
2	3	0	1
	4	2	3
3	5	0	1
	6	2	3

Figure 5.a shows a balanced incomplete block design, hereafter called BIB. The 6 possible pairs of treatments are randomly assigned so that each pair appears at one location. Each treatment appears at 3 locations. If the 250 subjects are divided equally between the treatments at each location, we will observe a total of 375 subjects per treatment – safely more that the 329 PROC POWER tells us to use. Figure 5.b. allows a direct comparison between the control treatment (0) and each experimental treatment in 2 locations instead of just one. If we allocate, say, 70 subjects to **treatment 0** at each location and the other 180 to the experimental treatment, we will have 420 total subjects assigned to **treatment 0** and 360 subjects assigned to each of the experimental treatments. Figure 5.c. shows a complete block design. This design requires us to disregard location distinctions in order to construct complete blocks. If we pair locations carefully, we *might* be able to form entities that we can defensibly call "blocks." We know, however, that researchers often do convenience blocking because it's easier or because their design

education ended with the randomized complete block design. No matter how carefully we form blocks, the complete block design will suffer increased within-block heterogeneity and we need our eyes wide open regarding the consequences.

We can use GLIMMIX to compute a power analysis for each design. This allows us 1) to determine if allocating the required number of subjects across the locations still yields acceptable power and 2) allows us to compare the designs. The latter is extremely important. In many planning exercises, the same sample size can yield strikingly different power characteristics when the subjects are allocated using different designs. This is why "power and sample size" is something of a misnomer – if done without regard to possibly more efficient design alternatives, power and sample size analysis completely misses the point of using statistics to inform research design.

The GLIMMIX-based power analysis occurs in 3 steps:

1. DATA step to create the exemplary data set
2. GLIMMIX step to compute the information needed to evaluate the non-central *F*
3. PROBABILITY EVALUATION step uses SAS® probability functions to determine power.

We now examine each step, starting with the BIB plan shown in Figure 5.a.

Step 1: create the exemplary data set:

```
data bib;
input loc @@;
n_subj=125;
 do g=1 to 2;
  input trt @@;
   do r=1 to 1; /* R = # multiples of locations used to increase power */
   location=(r-1)*6+loc;
   pi=0.15*(trt=0)+0.25*(trt=1)+0.30*(trt=2)+0.35*(trt=3);
   mu=n_subj*pi;
   output;
  end;
end;
datalines;
1  0 1
2  0 2
3  0 3
4  1 2
5  1 3
6  2 3
;
```

N_SUBJ assigns the number of subjects to be observed at each location for each treatment. DO R=1 to 1 seems superfluous, but it allows us to increase the number of locations in our power analysis if 6 locations proves inadequate (it does in this example – see below). Changing the statement to DO R=1 TO 2 would allow us to see if 12 locations is enough, DO R=1 TO 3 allows us to assess 18 locations, etc. Notice that increasing the design in multiples is the only way we can increase sample size and keep the design balanced. LOCATION gives each replicate of a given pair of treatments a unique identification. This will be important for the GLIMMIX step. PI defines the probabilities, π_i for each treatment. We set PI for **treatment 0** equal to 0.15 based on the information given above. We set PI for one of the experimental treatments – it doesn't matter which one – to 0.25 so we can assess power for the clinically relevant difference we said is of primary importance. We set PI for another experimental treatment – again it doesn't matter which one – to 0.35 to give us the worse-case power assessment we used above with PROC POWER. For the other experimental treatment, in the absence of another stated objective, it doesn't matter what we use for PI. Here we split the difference – this allows us to determine power for 0.15 vs 0.30 – we might want to know this in case it turns out to be prohibitively expensive to obtain adequate power for a 0.15 vs 0.25 difference. MU is the binomial expected value that we use in the GLIMMIX step. Notice that it is often a non-integer value. That's okay.

Step 2: obtain need statistics using GLIMMIX

```
proc glimmix data=bib initglm;
 class location trt;
 model mu/n_subj=trt;
 random intercept trt / subject=location;
 parms (0.10)(0.20)/hold=1,2;
```

```
/*random _residual_;
 parms (0.5)(1)(0.99)/hold=1,2,3;*/
 contrast 'c vs e1' trt 1 -1 0 0;
 contrast 'c vs e2' trt 1 0 -1 0;
 contrast 'c vs e3' trt 1 0 0 -1;
 contrast 'e1 v e3' trt 0 1 0 -1;
 ods output contrasts=f_tests_b;
```

The CLASS, MODEL and RANDOM statements are almost identical to what we eventually use to analyze the data once we complete the study. The only difference: here we use MU; later, we will use Y. Notice that the linear predictor here is identical to the linear predictor we used in Example 1; all of these designs are 4-treatment versions of the same structure: binomial data with a blocked design.

PARMS gives our best anticipation of the variance components for LOCATION and TRT by LOCATION. In this case, we set $\sigma^2_L = 0.10$ and $\sigma^2_{TL} = 0.20$. The HOLD option prevents GLIMMIX from executing its variance estimation routines, instead directing it to use the values we supply for all computations that involve the variance components. Occasionally, GLIMMIX will abort, and the following message appears in the SASLOG:

```
ERROR: Values given in PARMS statement are not feasible.
```

The error is spurious, but results from the GLIMMIX procedure's internal architecture. The INITGLM option in the PROC statement overrides GLIMMIX's default starting values, making the error less likely. If it happens anyway, you can include a `random _residual_` statement and vary the overdispersion parameter incrementally down from 1. In a true binomial GLMM the scale parameter (ϕ) equals 1, but sometimes when you HOLD all the covariance components, GLIMMIX's internal architecture needs ϕ to be slightly less in order to run. A little trial-and-error with ϕ will solve the problem with negligible impact of the power computations. GLIMMIX with SAS® V9.2 is much less prone to this error than the earlier SAS® V9.1 edition, but it does occasionally happen.

The CONTRAST statements shown here target proportion differences of 0.10, 0.15 and 0.20 relative to the control's $\pi_0 = 0.15$ and the proportion difference 0.25 vs 0.35. These capture the differences whose power we want to assess. CONTRAST statements should be tailored to the objectives of the study you are planning. The ODS OUTPUT statement creates a SAS data set to be used in step 3.

Step 3. Use functions FINV and PROBF to determine power for the various tests.

```
data power_bib;
 set f_tests_b;
 alpha=0.05;
 ncp=numdf*fvalue;
 f_crit=finv(1-alpha,numdf,dendf,0);
 power=1-probF(f_crit,numdf,dendf,ncp);
proc print data=power_bib;
```

F_CRIT defines the critical value. This is the "table value" we look up; if the observed *F*-value exceeds this number, we reject $H_0 : \tau_i = \tau_{i'}$. POWER uses the PROBF function to determine the probability under the non-central *F* defined by our set of treatment difference, variance component, design structure and sample size. PROC PRINT lists the resulting power calculations.

Now for the results. Output 7 shows power for the comparisons defined by the CONTRAST statements for the 6-location BIB.

Output 7. Power for 6-location Balanced Incomplete Block, Binomial data,125 subjects per Loc x Trt

Obs	Label	NumDF	DenDF	FValue	ProbF	alpha	ncp	f_crit	power
1	c vs e1	1	3	2.05	0.2474	0.05	2.05259	10.1280	**0.17526**
2	c vs e2	1	3	4.03	0.1385	0.05	4.02540	10.1280	**0.29018**
3	c vs e3	1	3	6.39	0.0855	0.05	6.39373	10.1280	**0.41494**
4	e1 v e3	1	3	1.22	0.3496	0.05	1.22242	10.1280	**0.12488**

The last column, POWER, gives the power for each test. These numbers should make us sit up and take notice. PROC POWER – indeed any power/sample-size calculation based on pre-GLMM theory – gave 329 as the required number of subjects per treatment to achieve 80% power for our *worst case* comparison – here defined by the contrast

E1 V E3. The GLIMMIX-based power assessment gives our *best-case* comparison, $\pi = 0.15$ vs. $\pi = 0.35$ a power of only slightly greater than 41%. What is going on?

The GLIMMIX-based analysis is not wrong. You can easily show via simulation that, given the differences we've defined and the variance assumptions we've made, the results we see here are spot-on accurate. What we see here is the result of conventional power and sample size methodology's failure to account for random variation among locations. The variance assumptions used in this example, by the way, are not unusual for multi-location binomial data. In this model, we interpret σ_L^2 as the variance in the log-odds among locations and σ_{TL}^2 as the variance in log-odds-ratios among locations. While π_i expresses a treatment's probability at a typical, average location, we know that all locations are not created equal: favorable outcomes are more likely at certain locations that others and certain treatments do better a certain locations. Power analysis needs to account for this. As you can see, failure to account for it has dramatic consequences.

How do the other designs compare? The power analyses for the other two designs require some changes in the SAS statements. For the control vs. experimental treatment design (hereafter called CVT), Figure 5.b, the revised statements to create the exemplary data set are

```
data cvt;
input loc @@;
 do g=1 to 2;
  input trt @@;
 do r=1 to 1;
   location=(r-1)*6+loc;
   n_subj=70*(trt=0)+180*(trt>0);
   pi=(0.15*(trt=0)+0.25*(trt=1)+0.30*(trt=2)+0.35*(trt=3));
   mu=n_subj*pi;
   output;
  end;
end;
datalines;
1  0 1
2  0 2
3  0 3
4  0 1
5  0 2
6  0 3
;
```

Notice that we now write N_SUBJ so that **treatment 0** receives 70 subjects per locations whereas the other treatments receive 180. The only other change entails replacing the BIB data set with the CVT data set after DATALINES.

For the complete block design, Figure 5.c, the exemplary data set and GLIMMIX statements are

```
data rcb;
 do i=1 to 4;
   input trt @@;   n_subj=125;
 do block=1 to 3;
   mu=n_subj*(0.15*(trt=0)+0.25*(trt=1)+0.30*(trt=2)+0.35*(trt=3));
   output;
  end;
end;
datalines;
0 1 2 3
;
proc glimmix data=rcb initglm;
 class block trt;
 model mu/n_subj=trt;
 random intercept trt / subject=block;
 parms (0.0534)(0.2524)/hold=1,2;
/*random _residual_;
 parms (0.5)(1)(0.99)/hold=1,2,3;*/
 contrast 'c vs e1' trt 1 -1 0 0;
 contrast 'c vs e2' trt 1 0 -1 0;
```

<Living with Generalized Linear Mixed Models>, continued

```
contrast 'c vs e3' trt 1 0 0 -1;
contrast 'e1 v e3' trt 0 1 0 -1;
ods output contrasts=f_tests_r;
```

Because we have equal allocation of subjects, N_SUBJ reverts to 125, the same as for the BIB. Note the variance components. Because the blocks disregard the natural block size, they will be more internally heterogeneous and correspondingly less different from one another. We must take this into account when we compare complete and incomplete block designs. Also, the variance components for the complete block design are among blocks, not locations. So we re-label them and adjust their magnitude accordingly.

Conventional power and sample size allows you to account for the change in the residual variance – which would correspond to σ^2_{TL} (or σ^2_{BT} for the complete block design) in this example – for Gaussian data only. However, conventional software does not account for efficiency gains from recovery of inter-block information – here, the BIB and CVT benefit; the complete block design does not. Furthermore, as we've seen, conventional software does not account for either variance when the data are non-Gaussian.

Output 8 shows the power results for the control-vs.-experimental and the complete block designs.

Output 8. Power for 6-location control-vs.-experimental and 3-block complete block designs

a. Control vs. experimental

Obs	Label	NumDF	DenDF	FValue	ProbF	alpha	ncp	f_crit	power
1	c vs e1	1	3	2.11	0.2426	0.05	2.10595	10.1280	**0.17847**
2	c vs e2	1	3	4.13	0.1349	0.05	4.13331	10.1280	**0.29620**
3	c vs e3	1	3	6.57	0.0830	0.05	6.56852	10.1280	**0.42350**
4	e1 v e3	1	3	0.76	0.4476	0.05	0.75966	10.1280	**0.09655**

b. Complete Block

Obs	Label	NumDF	DenDF	FValue	ProbF	alpha	ncp	f_crit	power
1	c vs e1	1	6	1.99	0.2082	0.05	1.98856	5.98738	**0.22219**
2	c vs e2	1	6	3.90	0.0957	0.05	3.89987	5.98738	**0.38309**
3	c vs e3	1	6	6.19	0.0472	0.05	6.19441	5.98738	**0.55015**
4	e1 v e3	1	6	1.18	0.3183	0.05	1.18423	5.98738	**0.15172**

Here, there appears to be more to be gained than lost by combining locations into complete blocks. In this case the LOCATION variance $\sigma^2_L = 0.10$ is relatively small, so the variance that most affects power, σ^2_{TL} increases from 0.2 to $\sigma^2_{BT} = 0.25$, where σ^2_{BT} denotes the complete block x treatment variance, as opposed to the location x treatment variance for the incomplete block designs. In this case, the increase is offset by the complete block design's greater inherent efficiency, *all other things being equal*.

If we have a larger variance among locations, say $\sigma^2_L = 0.40$, and σ^2_{TL} remains 0.2, then the variance components for the complete block design would be $\sigma^2_B = 0.161$ and $\sigma^2_{BL} = 0.449$. If the variance among locations is even larger, σ^2_{BL} is further increased for the complete block design, whereas σ^2_{TL} remains constant for the two incomplete block designs. For $\sigma^2_L = 0.8$, $\sigma^2_B = 0.307$ and $\sigma^2_{BL} = 0.709$.

Output 9 shows how increased variability among the locations affects power for these designs.

Output 9. Impact of increased σ_L^2 on power for BIB, Ctl vs Exp (CVT) and Complete Block (RCB) designs

a. $\sigma_L^2 = 0.40$

Obs	Label	power_bib	power_cvt	power_rcb
1	c vs e1	0.16051	0.15976	0.15392
2	c vs e2	0.26273	0.26156	0.25487
3	c vs e3	0.37586	0.37440	0.37096
4	e1 v e3	0.11537	0.08501	0.11013

b. $\sigma_L^2 = 0.80$

Obs	Label	power_bib	power_cvt	power_rcb
1	c vs e1	0.15575	0.15237	0.11789
2	c vs e2	0.25378	0.24773	0.18424
3	c vs e3	0.36292	0.35439	0.26334
4	e1 v e3	0.11231	0.08119	0.08896

We see that as location variance increases, the penalty for disregarding natural block size (locations) in favor of arbitrary complete blocks increases. When $\sigma_L^2 = 0.80$, power with the complete block design suffers noticeably.

Regardless of design and location variance, so far in this example none of these designs has power remotely approaching acceptable. For the best-case design we have seen so far, the complete block design with $\sigma_L^2 = 0.10$, power for the design's primary objective, the C VS E1 contrast, is 0.22. Using the GLIMMIX power programs we can change N_SUBJ to examine the impact of increasing the number of subjects per location-treatment combination. Although space does not permit showing the results here, you can easily verify that increasing the number of subjects has almost no impact on power. The only way to address the power problem for these designs is to increase the number of locations.

For the best-case, complete block, $\sigma_L^2 = 0.10$, 13 blocks (obtained using DO BLOCK=1 TO 13 when defining the exemplary data set) yields power=0.81 for the C VS E1 contrast. This means that under the most optimistic scenario, we need at least 13 blocks and hence 26 locations. Instead of the 329 subjects per treatment requirement we obtained using conventional power software, we actually need 13 x 125 = 1625 subjects per treatment. Actually, if we use 26 locations, we can reduce the number of subjects per treatment x block to 100, or 1300 subjects per treatment total, and still maintain 80% power.

The take home message here is that we cannot accurately assess power for studies like this without using GLMM-based power assessment methods. If we say a source of variation exists at *data* analysis time, we have to account for it at *power* analysis time. Otherwise, we risk results that can be misleading in the extreme. The price an unwary researcher would pay here would be to conduct a study that probably takes considerable effort and expense even for 6 locations, believing that power is 80% when in reality power is at best little more than 20%. This is what I call a catastrophically inaccurate assessment of power.

At the risk of belaboring a point, keep in mind that this is not an exotic, advanced scenario. Example 3 is only a little more complicated than a basic, introductory level, first-semester-stat-methods design. People who do studies that demand advanced, exotic statistical methods probably already use state-of-the-art statistical theory. Those who stand to benefit most from the methodology we are discussing are students and mid-level consumers of statistical methods, the ones we currently "protect" from these ideas. Colleagues tell me "students have enough trouble just learning 'the basics.' If you tried to teach them these 'advanced ideas,' it would just exacerbate the problem." Which brings us to the third "living with GLIMMIX" challenge. Are the examples we've been considering "advanced ideas"? What, in 2011, *is* "standard statistical practice?" What constitutes "the basics"?

THE THIRD CHALLENGE: WHAT DO WE TEACH THE STUDENTS?

This section came perilously close to being entitled, "What do we tell the children?" A colleague of mine once said during a teaching workshop, "Never consciously teach anything you know you will have to unteach later." I believe these words to be true – and relevant to this presentation.

Consider Example 1. If we poll those who teach introductory graduate level courses in statistical methods, and ask them for a list of topics they consider basic and essential and those they consider optional or "advanced," it is safe to say that the paired *t*-test, the analysis of variance and the normal approximation to the binomial would all rank high on the list of "basics." It is also safe to say that the GLMM would rank much lower on this list. Given what Example 1 tells us, if we follow through on these rankings, we will have created a perfect opportunity to teach something that will need to be untaught. If our approach to intro classes is "ANOVA learnable, other stuff too hard" and we teach "if N=100, it's safe to treat a binomial as if it's normal," then what do we think users are going to do when they encounter data like Example 1? Later, if they are lucky they will realize that they spent a great deal of time and effort learning an antiquated method; if they're unlucky they'll go blithely on, unaware of the disconnect between the analysis they are doing and the interpretation they are attaching to it.

Several years ago, at a now half-remembered ENAR talk, the speaker expressed the opinion that academic statisticians spend an inordinate amount of their effort on the last 5%. Arguably, Example 1 is guilty of this – in most cases the discrepancy between the marginal and conditional analyses of binomial may indeed be too small to make much difference. This case is harder to make for Example 2. A misspecified error term causing a grossly inflated type I error rate is not merely "the last 5%." It is a serious thing. Working through some version of an exercise like the "what would Fisher do?" process is the only way to avoid mistakes of this magnitude. Design courses do this to some extent with Gaussian data, but students and practitioners also need to understand implications unique to *non-Gaussian* data. Most will encounter both types of data in real life.

What about Example 3? No doubt our poll of stat methods instructors would identify a basic sample size formula high on the essentials list. Perhaps methodology underlying our initial PROC POWER assessment of required sample size might make the "basics" list, especially if students inhabit a culture where power analyses are mandatory for grant applications and dissertation proposals. The discrepancy between power determined via "the basics" and power determined as we did it in the previous section should give us pause.

The problems are related. How *do* we talk about design and analysis in 2011?

Shortly after I became department chair in 2001, our former chair, who had about as much love for faculty meetings as I did (none) told me about team building activities that the university's recreation center had started offering. An acquaintance had taken part in one and spoke very highly of it. We decided to give it a try. It sounded better than sitting in a stuffy room drinking too much coffee and putting too much heat and too little light into issues far too inconsequential.

The activities consisted of elaborate obstacle courses that an individual could not negotiate alone. A team had to work together. If they didn't, nobody got across. If they did, everyone made it. You didn't have to be athletic, but you did have to be a team participant. As we went through these exercises, we began to learn, among other things, about our own approach to solving novel problems. I discovered that I don't have much patience for pre-planning. Being analytic people, we would try to break down the problem, think through how are we going to do this, how are we going to do that? At some point I would lose it and say "Let's just go! Some bridges we're not going to figure out how to cross till we actually get to them." Then I'd get halfway out on a limb (or up the creek without a paddle might be more like it) and have a "Help, I'm stuck – what now?" moment as I gave my colleagues that deer-in-the-headlights look. But at least the ice was broken and once they were on board we would figure out how to get the rest of way through. My style seems to be leap before you look, then improvise – and trust your colleagues to have good ideas once they see the problem.

Why the digression? I don't pretend to know exactly how a contemporary statistical methods, design or introductory linear models course should look, but I do have an idea of some basic principles. Let's consider the data analysis world we inhabit. Table 2 is an elaboration of one Lock and DeVeaux (2007) presented at the USCOTS meetings. Their talk concerned undergraduate statistics, but their main idea applies. In graduate level courses, we say the building block of statistical analysis is < response variable = explanatory variable + error > . Table 2 portrays a sense of what this expression must cover. It is not intended to be exhaustive.

Table 2. Response and Explanatory Variables in the Contemporary World

Response Variable example distribution	**Explanatory Variable a.k.a. Model** **Fixed Effects** **Categorical**	 **Continuous**	**Random Effects** multi-site, split-plot	**Correlated Errors** time / space
Proportion discrete 2-cateegory: binomial discrete >2 category: multinomial continuous: beta	contingency tables			
Count Poisson negative binomial				
Continuous Gaussian (normal)	$\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$ ANOVA	 regression		
Time to Event exponential				

The gray-shaded cells represent most of the subject matter of introductory statistical methods sequences. If the sequence includes a design component, it may tiptoe lightly in the blue cell. Historically, the introductory linear models course has a narrower focus – just the cells containing $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$. Table 2 reveals one obvious problem: those who go on to use what they learn in their stat methods or linear model *sequence* (usually two courses) will encounter problems elsewhere on the matrix – not just in the highlighted cells. The other problem connects this table

with our examples: in learning the highlighted cells, students learn habits of mind that will prove counterproductive – even maladaptive – when applied to the other 12 or 13 cells on Table 2, much as they were in Examples 1, 2 and 3. This is the part where we have to unteach what we taught – and it's the real problem.

One non-answer to the problem is to ask students to take more statistics classes. Disciplines that are "consumers of statistical methods" are also becoming more complex. They need statistics, but our classes can't be allowed to crowd out essential courses in their major. As for our students, there comes a point where the quality of a graduate program is inversely proportional to the number of rules and requirements it has. Adding more classes is not the answer. Using the time we have with students more wisely is.

There are many excellent textbooks on methods, linear models and design of experiments. This is not intended to be a knock on them, but if you look through them, you get a sense of what might change. We see a heavy focus on analysis of variance and regression methods. Discussion of probability, what little of it there is, focuses on the normal distribution, use of *t* and *F* tables and, in linear models texts, sums of squares-driven quadratic forms. Occasionally we see passing mention of PROC MIXED, maybe a paragraph on generalized linear models, but no hint of any recognition of what we've talked about in this paper.

With analysis of variance, we spend a great deal of time teaching students to do the arithmetic of one-way and perhaps two-way ANOVA tables. At some point we say, "in real life, you won't actually *do* these calculations; you will use the computer" and they are taught

```
proc glm;
 class trt;
 model y=trt;
```

and how to read the output. How much more difficult is

```
proc glimmix;
 class trt;
 model y=trt;
```

How hard would it be if from day one we also tell student that we have different kinds of data? Some has a bell-shaped curve, but we also have Y/N proportions, counts, etc. If you have data whose histogram looks like a bell-shaped curve, you can do the above. But if it's Y successes out of N, you do this

```
proc glimmix;
 class trt;
 model y/n=trt;
```

If it's counts, you do this

```
proc glimmix;
 class trt;
 model y=trt/d=poisson;
```

and you keep your eye on the residual plots and the Pearson χ^2/df . And so forth.

It is entirely possible that learning the GLIMMIX-based approach to power analysis would actually help students grasp what is supposed to go on in statistical planning. It is entirely possible that by reducing it to a formula, instead of protecting students from something that is too hard, we actually wind up communicating to students that planning is trivial and unimportant, just something you get a black box to do.

Not surprisingly, some of my colleagues are very skeptical of this. They worry that this will make things worse, not better. If we merely substitute GLIMMIX recipes for GLM recipes without changing how we approach the background leading up to the software, my skeptical colleagues are absolutely right. There is a clear line from ANOVA table arithmetic to PROC GLM. To truly embrace the paradigm shift, we have to change the conversation from sums of squares to *maximum likelihood* and figure out an accessible language for doing this.

Quadratic forms are essential, but they do not have to be tied exclusively to sums of squares. Sums of squares have no meaning in GLMMs, but the test statistics for GLMM estimable functions are quadratic forms, too – and much more widely applicable. Much of the energy we currently spend (and ask students to spend) on ordinary least squares would be energy better spent on likelihood-based theory and methods.

CONCLUSION

The ideas in Examples 1, 2 and 3 are not new. They have been well-known in modeling circles for at least two decades. Although I am constantly surprised: either statisticians are very good at compartmentalizing – what they know perfectly well when they wear their theory hat ceases to exist when they have their practitioner hat on – or these ideas truly have not penetrated the world of stat methods instructors and consulting statisticians.

Either way, the knowledge has been out there. The difference now is that the software is also available. Anyone who can use PROC GLM can use PROC GLIMMIX. Now it's time for courses – and practice – to catch up. Neither PROC GLM nor PROC MIXED forced us to reconsider how we present the run up from fundamentals to software implementation. We could go on teaching stat methods, linear models – and design of experiments, for that matter – using the same old same old. With PROC GLIMMIX business-as-usual won't fly. I hope the examples presented earlier help make the case that we *need* to reconsider our approach. Not a tweak – a complete overhaul.

In thinking about the transition, another expression of my "don't teach what you'll have to unteach" colleague came to mind: "We're building an airplane and trying to fly it at the same time." A better metaphor might be the bicycle trail. Sums of squares are like training wheels. Given the computing technology available to Fisher and Yates – pencil and paper and not much more – training wheels made a lot of sense. But if you want to ride on the bicycle trail, it's better to learn to ride without the training wheels. We need to spend more time teaching researchers in allied disciplines and statisticians in training how to ride bicycles instead of continuing to emphasize the tricycle and training wheels.

REFERENCES

1. Fisher, R.A. and Mackenzie, W.A. (1923) Studies in Crop Variation II: The manurial response of different potato varieties. *Journal of Agricultural Science*. **13** pp. 311-320
2. Littell, R.C. (1980) Examples of GLM applications. *SAS Users' Group International: Proceedings of the Fifth Annual Conference.* Cary, NC: SAS Institute, Inc. pp. 208-214.
3. Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D and Schabenberger, O. (2006) *SAS for Mixed Models, 2nd ed*. Cary, NC: SAS Institute, Inc.
4. Lock, R. and DeVeaux, D, (2007) The Second Course in Statistics. *United States Conference on Teaching Statistics*. Available at http://www.causeweb.org/uscots/uscots07/program/breakout1.php
5. Milliken, G.A. and Johnson, D.E. (2009) *Analysis of Messy Data. Vol. 1. Designed Experiments. 2nd ed.* Boca Raton, FL: Chapman and Hall/CRC.
6. O'Brien, R.G., and Lohr, V.I. (1984) Power analysis for univariate linear models: the SAS system makes it easy. *SAS Users' Group International: Proceedings of the Ninth Annual Conference.* Cary, NC: SAS Institute, Inc. pp. 840-846.
7. SAS Institute, Inc. (2008) *SAS/STAT® 9.2 User's Guide.* Cary, NC: SAS Institute, Inc.
8. Schabenberger, O. (2008) Aspects of the Analysis of Split-Plot Experiments. *JSM 2008 Section on Physical and Engineering Sciences.* Abstract available: http://www.amstat.org/meetings/jsm/2008/onlineprogram/index.cfm
9. Speed, T. (2010) And ANOVA Thing. *IMS Bulletin*. **39**:4. p. 16.
10. Yates, F. (1935) Complex Experiments (with discussion), *Supplement to the Journal of the Royal Statistical Society* **2** pp. 181-247.
11. Young, L.J., Campbell, N.L. and Capuano, G.A. (1998) Analysis of Overdispersed Count Data from Single-Factor Experiments: A Comparative Study. *J. Agricultural, Biological and Environmental Statistics*. **4**:3 pp. 258-275

CONTACT INFORMATION

Walt Stroup
University of Nebraska
Department of Statistics
Lincoln, NE 68583-0963
(402) 472-1149
wstroup@unl.edu