Paper 11663-2016

PROC GLIMMIX as a Teaching and Planning Tool for Experiment Design

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

ABSTRACT

In our book, *SAS® for Mixed Models*, (Littell, et al., 2006), we write, "the majority of modeling problems are really design problems." Graduate students and even relatively experienced statistical consultants can find translating a study design into a useful model to be a challenge. Generalized linear mixed models (GLMMs) complicate this challenge, because they accommodate complex designs, complex error structures, and non-Gaussian data. This talk covers strategies proven to be effective in design and modeling courses and consulting sessions. In addition, GLMM methods can be extremely useful in planning experiments. This talk discusses methods to implement precision and power analysis to help choose between competing designs and to assess the adequacy of proposed designs.

INTRODUCTION

The first edition of *SAS® for Mixed Models* (Littell, et al.) appeared in 1996, the 2nd edition in 2006, and the 3rd edition is currently a work in progress. Since its appearance, the authors have received numerous requests for help with modeling problems. A common theme of these requests, as well as courses we teach in statistical modeling, is that the majority of "modeling" issues have a strong component of design problems embedded in them. This paper addresses two interrelated aspects of this reality. First, conventional power and sample size algorithms are not up to the task of providing the needed information to plan studies when the design complexity calls for mixed model analysis. This is especially true when the primary response variable does not have a Gaussian (a.k.a. "normal") distribution. As a result, designs that fit this description often have serious shortcomings. Second, graduate students, both in statistics and in fields that are consumers of statistical methods, have trouble translating descriptions of study designs into reasonable models. Methods taught to them are often antiquated and do not address the realities of generalized and mixed models. Even experienced statistical consultants often struggle for the same reason.

This paper presents two sets of tools that use PROC GLIMMIX. One concerns generalized linear mixed model (GLMM) based power and precision analysis. The other concerns a process to help translate study design into plausible models. Both tools are useful in teaching linear models and design of experiments courses, as well as for statistical consulting, and for planning research about GLMMs.

Part 1 of this paper reviews GLMM-based power and precision analysis, and Stroup's WWFD ("What Would Fisher Do?") process for translating design into model. Part 2 presents an introductory how-to illustration using mixed model tools to choose among competing blocked designs. Part 3 presents examples of the WWFD process to determine appropriate models for blocked designs with non-Gaussian data. Part 4 extends the precision and power methods presented in Part 2 to non-Gaussian data. Finally, Part 5 presents an application of the WWFD process to a design structure whose complexity proved to be daunting even to statisticians with extensive consulting experience.

PART ONE – GLMM-BASED DESIGN PLANNING AND MODEL CONSTRUCTION BASICS

Design Planning

Power analysis using SAS linear model software was introduced by Littell (1980, Lohr and O'Brien (1984) and O'Brien and Lohr (1984). Their approach was adapted for mixed models by Stroup (1999) and Littell, et al. (2006). Stroup (2011, 2013) presented a further extension to the GLMM. The following summarizes the essentials.

The GLMM is defined by the following elements. Start with a vector of observations whose distribution, conditional on random model effects, can be written $\boldsymbol{y}|\boldsymbol{b}\sim\mathcal{D}\left(\boldsymbol{\mu}|\boldsymbol{b},\boldsymbol{V}_{\mu}^{1/2}\boldsymbol{P}\boldsymbol{V}_{\mu}^{1/2}\right)$, where $\mathcal{D}$ denotes either a distribution that belongs to the exponential family, or a validly defined quasi-likelihood, $\boldsymbol{\mu}|\boldsymbol{b}$ denotes the expected value of $\boldsymbol{y}$ given $\boldsymbol{b}$, $\boldsymbol{V}_{\mu}^{1/2} = diag\left[\sqrt{v(\mu)}\right]$ where $v(\mu)$ denotes the variance function for $\mathcal{D}$, and $\boldsymbol{P}$ denotes a covariance (if the distribution is Gaussian), scale matrix (for non-Gaussian distributions) or working covariance (if $\mathcal{D}$ is a quasi-likelihood). The vector of random effects is assumed to have a Gaussian distribution, written $\boldsymbol{b}\sim N(\boldsymbol{0},\boldsymbol{G})$. The GLMM fits a linear predictor $\boldsymbol{X\beta}+\boldsymbol{Zb}$ to a link function $\boldsymbol{\eta}=g(\boldsymbol{\mu}|\boldsymbol{b})$. On default, the GLIMMIX procedure fits the model using pseudo-likelihood (PL) estimating equations $\begin{bmatrix}\boldsymbol{X'WX} & \boldsymbol{X'WZ}\\ \boldsymbol{Z'WX} & \boldsymbol{Z'WZ}+\boldsymbol{G}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{b}\end{bmatrix}=\begin{bmatrix}\boldsymbol{X'Wy}^{*}\\ \boldsymbol{Z'Wy}^{*}\end{bmatrix}$ where $\boldsymbol{W}=\left[\boldsymbol{D}\left(\boldsymbol{V}_{\mu}^{1/2}\boldsymbol{P}\boldsymbol{V}_{\mu}^{1/2}\right)\boldsymbol{D}\right]^{-1}$, $\boldsymbol{D}=diag[\partial\boldsymbol{\eta}/\partial\boldsymbol{\mu}]$, and $\boldsymbol{y}^{*}=\boldsymbol{\eta}+\boldsymbol{D}(\boldsymbol{y}-\boldsymbol{\mu})$. Once the model is fit, inference with GLIMMIX depends on assuming that the vector of

prediction errors, $\begin{bmatrix}\hat{\beta}-\beta \\ \hat{b}-b\end{bmatrix}$ is approximately distributed $N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \boldsymbol{C}\right)$, where $\boldsymbol{C}$ denotes the generalized inverse of $\begin{bmatrix}\boldsymbol{X'WX} & \boldsymbol{X'WZ} \\ \boldsymbol{Z'WX} & \boldsymbol{Z'WZ}+\boldsymbol{G^{-1}}\end{bmatrix}$.

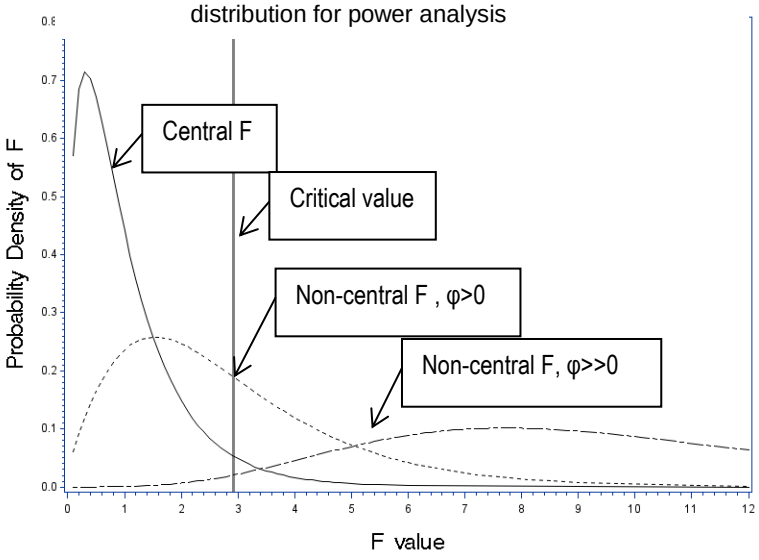
Of specific interest for power and precision analysis is the above result's application to estimable functions $\boldsymbol{K'\beta}$. First, the variance of $\boldsymbol{K'\hat{\beta}}$ is $Var\left(\boldsymbol{K'\hat{\beta}}\right)=\boldsymbol{K'}\left(\boldsymbol{X'}(\boldsymbol{V}^*)^{-\mathbf{1}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{K}$, where $\boldsymbol{V}^*=\boldsymbol{ZGZ'}+\boldsymbol{R}^*$ and $\boldsymbol{R}^*=\boldsymbol{D}\left(\boldsymbol{V_\mu^{1/2}PV_\mu^{1/2}}\right)\boldsymbol{D}$. Second, the Wald statistic divided by the rank of $\boldsymbol{K}$, $\left(\boldsymbol{K'\hat{\beta}}\right)'\left[\boldsymbol{K'}\left(\boldsymbol{X'}(\boldsymbol{V}^*)^{-\mathbf{1}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{K}\right]^{-1}\left(\boldsymbol{K'\hat{\beta}}\right)/rank(\boldsymbol{K})$, referred to as the approximate *F*-statistic, is approximately distributed $F_{(rank(\boldsymbol{K}),\nu,\varphi)}$ where ν denotes the denominator degrees of freedom involved in estimating $Var\left(\boldsymbol{K'\hat{\beta}}\right)$ and $\varphi=(\boldsymbol{K'\beta})'\left[\boldsymbol{K'}\left(\boldsymbol{X'}(\boldsymbol{V}^*)^{-\mathbf{1}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{K}\right]^{-1}(\boldsymbol{K'\beta})$ is the non-centrality parameter.

These results can be used for precision or power analysis of a proposed study design. First, let's be clear about what these are. **Precision analysis** refers to determining the expected standard error, or width of a confidence interval, for specified contrasts – i.e. $\boldsymbol{K'\beta}$ – that address the objectives of the study. **Power analysis** refers to determining the probability of rejecting H_0: $\boldsymbol{K'\beta}=\mathbf{0}$ for an estimable function of interest when $\boldsymbol{K'\beta}$ differs from zero by a specified amount. **Precision and power analysis** both require a proposed study design, including the number of proposed replications (a.k.a. "sample size"), and the components of variance associated with the ***G*** and ***R**** matrices. The proposed study design defines the ***X*** and ***Z*** matrices. **Power analysis** also requires one to specify the amount by which $\boldsymbol{K'\beta}$ is expected to differ from zero (a.k.a. the "effect size"). Usually, this is framed in terms of the minimum difference that would be considered scientifically, practically, or economically important – or, to borrow a phrase from the pharmaceutical world, the "minimum clinically relevant difference."

The following steps implement precision and power analysis with SAS software:

1. Create a data set identical to the one you will eventually use to analyze the data when it is collected, except that instead of actual data – which you do not yet have – use anticipated expected values for each treatment or treatment combination. Following Lohr and O"Brien, we refer to this as the **exemplary data set**. The means you use in the exemplary data set are not intrinsically important, but it is important that differences between means of treatments you intend to compare reflect the differences you consider to be scientifically relevant.
2. Use SAS mixed model software to compute the anticipated standard errors and test statistics. If your response variable is Gaussian, you can use either PROC GLIMMIX or PROC MIXED. If your response variable is non-Gaussian, you must use PROC GLIMMIX. In this step, you must provide values for the variance of all random model effects (the components of ***G***) and all residual covariance or working covariance terms (the components of ***R****).
3. For precision analysis, the standard errors from the GLIMMIX or MIXED step provide the needed information. In addition, confidence interval widths can be obtained if desired.
4. For power analysis, use the *F*-value from the GLIMMIX or MIXED step. You can see from the definitions of the approximate *F* statistic and the non-centrality parameter given above, that multiplying the GLIMMIX/MIXED *F*-value by the corresponding numerator degrees of freedom gives you the non-centrality parameter. You then use SAS probability functions to compute power. The idea is illustrated in Figure 1.

Figure 1. Illustration of Central and non-Central F distribution for power analysis



Under the null hypothesis, the treatment effect $\boldsymbol{K'\beta}=\mathbf{0}$. Hence the non-centrality parameter is 0, and the *F* statistic has an approximate central *F* distribution. The critical value is the value of the central *F* such that the area under the curve to the right is α, the probability of a type I error, i.e. of rejecting H_0 when H_0 is true. When the treatment effect is non-zero, the non-centrality parameter is greater than zero, shifting the distribution of the *F* statistic to the right. The greater the non-centrality parameter (i.e. the greater the treatment effect size) the more the distribution shifts

to the right. Because you reject H_0 if the *F* statistic is greater than the critical value, the area under the curve to the right of the critical value gives the probability of rejecting H_0 when H_0 is false. That is, the area to the right of the critical value for non-central *F* is the power for the corresponding treatment effect size.

Examples demonstrating the implementation of precision and power analysis, and why GLMM-based methods are critically different from most conventional, commercially available power and sample size software, are shown in subsequent sections.

Model Construction

Littell, et al. (2006) describe statistical models as "...mathematical descriptions of how data conceivably can be produced." In practice, models serve as a template for data analysis. The two aspects of a model are *how the data* arose and the *template*. Beginning students and unsophisticated practitioners of statistics tend focus exclusively on the *template*, and pay little attention to the arguably more important *how data arose* narrative. Disregarding the latter frequently results in a disconnect between the study design and the model, and is among the leading causes of modeling problems. Stroup (2011, 2013) created WWFD (What Would Fisher Do?) as a teaching tool intended to help linear models students (and statistical consultants) address this problem.

WWFD has its roots in strategies employed for messy data by Milliken and Johnson (2009), but was directly inspired by Fisher's comments following Yates (1935) paper "Complex Experiments." Fisher said that any study could be characterized in terms of its "topographical" and "treatment" aspects. Federer (1955) and Milliken and Johnson use the terms "experiment design" and the "treatment design." WWFD starts with a loose interpretation of Fisher. Write separate ANOVA sources of variation and degrees of freedom for each aspect and then combined them. Once you do this, add a GLMM twist.

To illustrate, consider a randomized block design. The "topographical," a.k.a. "experiment design" ANOVA can be written

Source of Variation	d.f.
block	$b-1$
exp. units (block)	$b(u-1)$
TOTAL	$bu-1$

where *b* denotes the number of blocks and *u* denotes the number of experimental units per block. Here we assume that each block has the same number of experimental units. Adding the treatment aspect (assuming *t* treatments) yields

"Topographical"		**"Treatment"**	
Source of Variation	d.f.	Source of Variation	d.f.
block	$b-1$		
		treatment	$t-1$
exp. units (block)	$b(u-1)$	"parallels"	$bu-t$
TOTAL	$bu-1$	TOTAL	$bu-1$

Note that this process does not necessarily assume a complete block design. The position of the treatment source of variation matters: it should be placed in the line above the unit to which treatment levels are randomly assigned. The term "parallels" is Fisher's. He used the term to mean all the leftover sources of variation after accounting for treatment in the "treatment" ANOVA. "Parallels" play no role in the combined ANOVA, but placing it in the treatment ANOVA helps keep d.f. bookkeeping accurate.

The combined ANOVA appears in Table 1.

Table 1. Summary of WWFD ANOVA for Blocked Designs

"Topographical"		**"Treatment"**		**Combined**	
Source	d.f.	Source	d.f.	Source of Variation	d.f.
block	$b-1$			block	$b-1$
		treatment	$t-1$	treatment	$t-1$
exp. units (block)	$b(u-1)$	"parallels"	$bu-t$	e.u. (block) | trt a.k.a. block × trt a.k.a. "residual"	$bu-b-t+1$
TOTAL	$bu-1$	TOTAL	$bu-1$	TOTAL	$bu-1$

Traditionally, linear models students are taught to write the model as a one-to-one translation of the combined ANOVA into the equation $y_{ij} = \mu + b_j + \tau_i + e_{ij}$. That is, $\text{block} \Rightarrow b_j$, $\text{treatment} \Rightarrow \tau_i$ and e.u.(block) | trt (my notation for "unit within block after accounting for treatment") usually called "error" or "residual" in most textbooks $\Rightarrow e_{ij}$. GLMM textbooks call this the "model equation approach." It works – usually – if the data have a Gaussian distribution, but it obstructs constructing a plausible model if the data are not Gaussian.

To use the combined ANOVA to construct a model consistent with the study design, start with the unit on which observations are taken, in this case, e.u.(block) | trt. For a properly constructed combined ANOVA, the unit of observation will always correspond to the source of variation in the last line. Write the probability distribution considered plausible for observations at the unit level. For example, if independent, homoscedastic Gaussian observations are assumed, $y_{ij} \sim NI\left(\mu_{ij}, \sigma^2\right)$, where NI denotes "normal and independent." On the other hand, if the observations are N independent binary observations per experimental unit, and the response of interest is the number of successes out of N, then a plausible distribution is $y_{ij} \sim \text{Binomial}\left(N, \pi_{ij}\right)$ where π_{ij} denotes the probability of a success for the j^{th} block, i^{th} treatment.

The next model-construction step involves deciding how the other sources of variation affect the expected value of the observations. In GLMM language, this means choose the link function and the linear predictor. For Gaussian data, the standard link is the identity and the standard linear predictor for a randomized block design is $\mu_{ij} = \mu + b_j + \tau_i$. For binomial data, the standard link is the logit, $\text{logit}\left(\pi_{ij}\right) = \log\left[\pi_{ij} / \left(1 - \pi_{ij}\right)\right] = \eta_{ij}$. If one follows precedent, the linear predictor would be $\eta_{ij} = \eta + b_j + \tau_i$. However, in GLMM terms, this is called the "naive" linear predictor. The reason is that unit(block)|trt is itself a source of variation. That is, individual units have variability over and above what is accounted for by block and treatment. The Gaussian model accounts for this by requiring an estimate of σ^2 in addition to the linear predictor. However, with the binomial model, once you estimate the linear predictor, you have the estimated mean, proportional to $\hat{\pi}_{ij}$ and the variance, proportional to $\hat{\pi}_{ij}\left(1 - \hat{\pi}_{ij}\right)$. There is no additional variance to estimate. This leads to *overdispersion* – failure of the model to account for all of the variability in the data. There are several strategies to deal with this, some of which are presented below. The point here is that paying attention to what the combined ANOVA is telling you about how the data arose, and not merely using it to compute sums of square or write a model equation, helps construct a sensible model.

The final WWFD step involves deciding which, in any, of the effects in the linear predictor have a probability distribution. For example, if the blocks in the study represent a sample of the target population, and one intends to infer results of the study to that population, then, by definition, the block effect should have a probability distribution. This leads to the final step of WWFD. Write the combined ANOVA, the assumed distribution of the observations, the effects that will account for sources of variation, and the assumed distribution of effects to be regarded as random. Table 2 shows an example of the result.

Table 2. Model Construction Steps for Gaussian and Binomial Data using Block Design ANOVA

Combined		Model Effects Data Gaussian	Model Effects Data Binomial
Source of Variation	d.f.		
block	$b-1$	$b_j \sim NI\left(0, \sigma_B^2\right)$	$b_j \sim NI\left(0, \sigma_B^2\right)$
treatment	$t-1$	τ_i	τ_i
e.u. (block) | trt a.k.a. block × trt a.k.a. "residual"	$bu-b-t+1$	observation at e.u. level $y_{ij} \mid b_j \sim NI\left(\mu_{ij}, \sigma^2\right)$ $\mu_{ij} = \mu + \tau_i + b_j$	additional term at e.u. level needed (see Example 2) observation at e.u. level $y_{ij} \mid b_j \sim Binomial\left(N_{ij}, \pi_{ij}\right)$ naive linear predictor $\eta_{ij} = g\left(\pi_{ij}\right) = \eta + \tau_i + b_j$ naivety addressed Example 2

In principle, WWFD can be applied to designs of arbitrary complexity. The example in Part 5 will illustrate the process for a complex design.

To summarize, the steps of the WWFD ANOVA process are

- obtain the combined ANOVA sources of variation from the "topographical" and "treatment" ANOVAs

- write the assumed unit-level distribution, i.e. the distribution of the observations conditional on sources of variation assumed to be random. In GLMM theory **y** denotes the observation vector, **b** denotes the random effect vector, and $f(\mathbf{y}|\mathbf{b})$ denotes the p.d.f. of the unit-level distribution.
- write the link function $\boldsymbol{\eta} = g(\boldsymbol{\mu}|\mathbf{b})$, where $\boldsymbol{\mu}|\mathbf{b} = E(\mathbf{y}|\mathbf{b})$
- write the linear predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$
- write the assumed distribution of those effects considered random. The p.d.f. of the random effects is denoted $f(\mathbf{b})$.

PART TWO – BLOCKED DESIGNS I: PRECISION ANALYSIS WITH GAUSSIAN DATA

The following examples can be used to illustrate precision and power analysis with blocked designs. In a design of experiments class, the blocked design can be viewed narrowly as randomized complete block and incomplete block designs. Viewed more broadly, they include paired or grouped comparisons, matched pairs or groups, stratified or cluster sampling designs, etc. The first example shows how to use GLMM-based precision and power analysis to choose among competing designs that have identical sample size but very different precision and power characteristics. This example illustrates a key concept often lost on unsophisticated practitioners: design structure matters – sample size alone does not provide a complete picture.

Example 1: choosing among competing designs. Suppose that we have 10 blocks, each of size 3. These could be 10 schools participating in a study to compare different curricula, with 3 classrooms per school available, or 10 benches in a greenhouse, with each bench having room for 3 treatments, or 10 clinics each with the capacity to handle 3 groups of patients. Suppose that the objectives of the study require comparing 6 treatments. Figure 2 shows three plausible designs.

Figure 2. Possible Designs for 10 block, 6 treatment study with block size 3

Balanced Incomplete Block

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

Control (trt 1) vs. All Others

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

Disconnected*

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

* more properly, group-nested or split-plot

Following the WWFD process in the previous section, the linear predictor for these designs can be written $\eta_{ij} = \eta + \tau_i + b_j$ with the block effects assumed random and distributed $NI(0, \sigma_B^2)$. The following SAS statements (next page) allow you to compare the precision of these designs:

Step 1. Create Exemplary Data Set

```
/* balanced incomplete block */
   data bib;
    input block @@;
     do eu=1 to 3;
      input trt @@;
      mu=0;
      output;
     end;
   datalines;
    1   1 2 3
    2   1 2 4
    3   1 3 5
    4   1 4 5
    5   1 5 6
    6   2 3 6
    7   2 4 5
    8   2 5 6
    9   3 4 5
   10   3 4 6
   ;
```

You can modify the above to create the exemplary data set for the Control vs. All Others and Group-Nested/Split-Plot designs. For precision analysis, it does not matter what value of the response variable you enter, because all you are interested in is the standard error of treatment mean differences and the expected width of confidence intervals. In the above program, the response variable is MU=0. In my linear models class at the University of Nebraska, one of my students defined the response variable as GO_BIG_RED=42. Next is the GLIMMIX step. Here, you must specify the assumed distribution of the response variable and the variance components. On default, GLIMMIX treats response variables as Gaussian. Let's start there, then modify the example for binomial and count data.

At this point in the example, we are interested in the relative precision of each design. This will allow you to choose which design best suits your needs. Then you can do more detailed assessment of the design you choose. To obtain relative precision, run step two.

Step 2. GLIMMIX Program for Precision Analysis

```
   proc glimmix data=bib;
    class block trt;
    model mu=trt;
    random intercept / subject=block;
    parms (0.5)(1)/hold=1,2;
    lsmeans trt/diff cl;
   run;
```

The PARMS statement specifies the block and residual variance, respectively. HOLD=1,2 instructs GLIMMIX to use the numbers given in the PARMS statement and hold them fixed. Otherwise, GLIMMIX would attempt to estimate variance from data with no variation, and you would get an error message. If the actual residual variance is σ^2 and you specify residual variance = 1, the standard errors in the DIFF listing from the LSMEANS statement are the multiples of the actual residual standard deviation. That is, the expected standard error of a treatment mean difference is <std error in DIFF listing>× σ. Because the residual variance is identical for all three designs under consideration, σ will be the same, so the standard error in the DIFF listing is all you need to compare the designs. Notice that the block variance in this program is actually the ratio σ_B^2/σ^2. This requires some knowledge. You need to elicit this information from the subject matter expert. While the ratio does not have a huge impact, it does affect recovery of inter-block information – the smaller the ratio, the greater the impact of inter-block information – and it has a huge impact in the next part of this example, where we compare these three designs to a complete block alternative.

The listing for the BIB example appears in Output 1.

Output 1. Listing for Precision analysis of BIB design with Gaussian Data

Differences of trt Least Squares Means									
trt	_trt	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
1	2	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
1	3	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
1	4	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
1	5	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
1	6	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
2	3	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
2	4	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
2	5	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
2	6	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
3	4	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
3	5	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
3	6	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
4	5	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
4	6	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370
5	6	0	0.6742	15	0.00	1.0000	0.05	-1.4370	1.4370

The items of primary interest are the Standard Error and the Lower and Upper confidence limits. These give the expected standard error of a treatment mean difference and the expected width of a 95% confidence interval for the estimate of a difference for each pair of treatments. You can see that the standard errors are the same for all treatment differences – hence why the design is called a "*balanced* incomplete block" – and the expected standard error is $0.67 \times \sigma$. Once you run the suitably modified program for the other two designs, you can merge the results into tables along the lines of Output 2.

Output 2. Standard error of a difference and expected Confidence Limits for BIB, CVO and SPLT designs

Obs	trt	_trt	bib_sed	cvo_sed	splt_sed
1	1	2	0.67420	0.62072	0.63246
2	1	3	0.67420	0.62072	0.63246
3	1	4	0.67420	0.62072	0.77460
4	1	5	0.67420	0.62072	0.77460
5	1	6	0.67420	0.62072	0.77460
6	2	3	0.67420	0.76696	0.63246
7	2	4	0.67420	0.76696	0.77460
8	2	5	0.67420	0.76696	0.77460
9	2	6	0.67420	0.76696	0.77460
10	3	4	0.67420	0.76696	0.77460
11	3	5	0.67420	0.76696	0.77460
12	3	6	0.67420	0.76696	0.77460
13	4	5	0.67420	0.76696	0.63246
14	4	6	0.67420	0.76696	0.63246
15	5	6	0.67420	0.76696	0.63246

Obs	trt	_trt	bib_LCL	bib_UCL	cvo_LCL	cvo_UCL	splt_LCL	splt_UCL
1	1	2	-1.43702	1.43702	-1.32303	1.32303	-1.34075	1.34075
2	1	3	-1.43702	1.43702	-1.32303	1.32303	-1.34075	1.34075
3	1	4	-1.43702	1.43702	-1.32303	1.32303	-1.64207	1.64207
4	1	5	-1.43702	1.43702	-1.32303	1.32303	-1.64207	1.64207
5	1	6	-1.43702	1.43702	-1.32303	1.32303	-1.64207	1.64207
6	2	3	-1.43702	1.43702	-1.63475	1.63475	-1.34075	1.34075
7	2	4	-1.43702	1.43702	-1.63475	1.63475	-1.64207	1.64207
8	2	5	-1.43702	1.43702	-1.63475	1.63475	-1.64207	1.64207
9	2	6	-1.43702	1.43702	-1.63475	1.63475	-1.64207	1.64207
10	3	4	-1.43702	1.43702	-1.63475	1.63475	-1.64207	1.64207
11	3	5	-1.43702	1.43702	-1.63475	1.63475	-1.64207	1.64207
12	3	6	-1.43702	1.43702	-1.63475	1.63475	-1.64207	1.64207
13	4	5	-1.43702	1.43702	-1.63475	1.63475	-1.34075	1.34075
14	4	6	-1.43702	1.43702	-1.63475	1.63475	-1.34075	1.34075
15	5	6	-1.43702	1.43702	-1.63475	1.63475	-1.34075	1.34075

Interpretation: The three columns whose labels end with _SED are the standard error of a treatment mean difference for the three designs: BIB refers to “balanced incomplete block,” CVO to “control vs. all others,” and SPLT to the disconnected/group-nested/split-plot design. The LCL and UCL columns refer to the lower and upper confidence bounds. The CVO design provides greater precision and narrower confidence intervals for differences between the control (treatment 1) and each other treatment, but comparisons between pairs of non-control treatments (2 through 6) are less precise and have wider confidence intervals. The SPLT design provides greater precision for differences within each group – {1,2,3} and {4,5,6} – but much less precision for comparisons across groups. This allows researchers to make informed decisions about which design to use.

One additional design to consider would be a complete block design. Notice that combining blocks {1,2}, {3,4}, {5,6}, {7,8} and (9,10} in the disconnected/group-nested/split-plot design yields a complete block design with equivalent replication per treatment. Many design of experiments textbooks give efficiency factors, i.e. the loss in precision using an incomplete block design instead of a complete block design. You can determine the efficiency factor for the BIB using the standard error of a difference computed above for the BIB, and the following statements to obtain the standard error of a treatment difference for a 5 block, 6 treatment complete block design:

```
proc glimmix data=rcbd;
 class block trt;
 model mu=trt;
 random intercept / subject=block;
 parms (0.5)(1)/hold=1,2;
 lsmeans trt/diff cl;
run;
```

Selected results appear in Output 3. Only the line for TRT 1 versus 2 appears, because, as with the BIB, for a complete block design, all difference standard errors are equal

Output 3. Difference Listing for 5-block, 6-treatment Complete Block Design

Differences of trt Least Squares Means									
trt	_trt	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
1	2	0	0.6325	20	0.00	1.0000	0.05	-1.3193	1.3193

The efficiency factor is $(0.6742/0.6325)^2 = 1.136$, interpreted as a 13.6% loss of efficiency – variance of an estimated treatment difference is 13.6% greater – if you use a BIB instead of complete block design. However, this efficiency factor is misleading. It assumes experimental unit variance is unchanged. In reality, larger blocks tend to increase within-block heterogeneity, meaning σ^2 does not remain constant. In this example, two natural blocks of size 3 must be combined to construct complete blocks of size 6. Experimental unit variance *will* increase as a result – the question is how much? Stroup (2013) shows an exact analytic way to determine the impact of larger blocks on σ^2 but it is easier to approximate this result by simulation. Use the following statements.

```
/* incomplete to complete block variance converter */
/* step one - block and e.u. variation             */
/*  assuming natural block size                    */
data natural_block;
seed=51722205;
block_var=0.5;
eu_var=1;
mu=0;
N_Natural_blocks=10;
natural_block_size=3;
do expt=1 to 1000;
 do natural_blk=1 to N_natural_blocks;
  blk_effect=sqrt(block_var)*rannor(seed);
  do eu=1 to Natural_Block_Size;
   e_ij=sqrt(eu_var)*rannor(seed);
   y=mu+blk_effect+e_ij;
   output;
  end;
 end;
end;
proc sort data=natural_variation;
 by expt natural_blk eu;
run;
/* step 2 - match e.u. to block they will be in    */
/*  under incomplete and complete block design     */
/* also give trt under complete block              */
/* so variance estimation can account for trt d.f. */
data create_rcb;
 input natural_blk complete_blk @@;
   do eu=1 to 3;
    input trt @@;
     do expt=1 to 1000;
     output;
     end;
   end;
datalines;
 1  1  1 2 3
 2  1  4 5 6
 3  2  1 2 3
 4  2  4 5 6
 5  3  1 2 3
 6  3  4 5 6
 7  4  1 2 3
 8  4  4 5 6
 9  5  1 2 3
10  5  4 5 6
;
proc sort data=create_rcb;
 by expt natural_blk eu;
run;/* merge data sets */
data combined;
 merge natural_block create_rcb;
 by expt natural_blk eu;
/* determine block and residual variance */
/*   for complete block design           */
```

```
ods results off;
ods html exclude all;
proc glimmix data=combined;
 by expt;
 class complete_blk trt;
 model y=trt ;
 random intercept / subject=complete_blk;
 ods output covparms=cb_var_est;
run;
proc sort data=cb_var_est;
 by covparm;
ods results on;
ods html select all;
proc means data=cb_var_est;
 by covparm;
 var estimate;
run;
```

Output 4 shows the results.

Output 4. Block and residual variance when complete blocks are constructed from smaller natural-size blocks

Parameter=Intercept (Block Variance)

N	Mean	Std Dev	Minimum	Maximum
1000	0.2150253	0.2563148	0	1.6253704

Parameter=Residual (Exp. Unit Variance)

N	Mean	Std Dev	Minimum	Maximum
1000	1.2863951	0.4524624	0.3536870	3.2288534

Given the block and experimental unit variance for naturally occurring blocks of size 3, constructing blocks of size 6 alters the block variance from 0.5 to 0.22 and the experimental unit variance from 1.0 to 1.29. The appropriate GLIMMIX step for the complete block design therefore should be

```
proc glimmix data=rcbd;
 class block trt;
 model mu=trt;
 random intercept / subject=block;
 parms (0.22)(1.29)/hold=1,2;
 lsmeans trt/diff cl;
run;
```

This yields a standard error of a difference of 0.7183, and an efficiency of $(0.6742/0.7128)^2 = 0.8462$. Contrary to the textbook efficiency factor, the complete block is actually just over 15% *less* efficient than the BIB. Note that the result is specific to a σ_B^2/σ^2 ratio of 0.5. The loss of efficiency increases as the ratio increases. The only case for which the textbook efficiency factor is correct is when the ratio is zero – in which case, why block at all?

You can use the exemplary data set / GLIMMIX step method for precision analysis when the primary response variable is non-Gaussian, or you can add a third step to assess power. These will be demonstrated below. Before doing precision analysis for non-Gaussian data, we need to address the "naive GLMM issue" raised in the WWFD introduction in Part One.

PART THREE - NAIVE, SENSIBLE AND NONSENSE MODELS FOR NON-GAUSSIAN DATA

In Part One we presented the WWFD ANOVA table for binomial data and noted that borrowing the linear predictor from the Gaussian LMM produced a naive model. The reason is that the estimated linear predictor, $\hat{\eta}_{ij} = \hat{\eta} + \hat{\tau}_i + \hat{b}_j$ accounts for the treatment and block sources of variation, but does not account for unit-level uniqueness. In a Gaussian model, the linear predictor gives you $\hat{\mu}_{ij}$ but you still must estimate the unit-level variance, σ^2. With binomial data, the linear predictor yields $\hat{\pi}_{ij} = 1/\left(1+e^{-\hat{\eta}_{ij}}\right)$, giving you both the mean and the variance – the latter being a function of $\pi_{ij}\left(1-\pi_{ij}\right)$. No separate estimate of the unit-level variance is needed if – and this is a big if – the model

Table 3. WWFD ANOVA and Sensible Models

Source of Variation	Gaussian LMM	Naive GLMM	GLMM with Gaussian Unit Effect	CS GEE	Beta-Binomial	Poisson-Gamma = Negative Binomial	Beta Binomial or Neg Binomial GEE	BetaBinomial or NegBinomial with unit level random effect
block	$b_j \sim NI\left(0,\sigma_B^2\right)$	$b_j \sim NI\left(0,\sigma_B^2\right)$	$b_j \sim NI\left(0,\sigma_B^2\right)$	σ_{CS}	$b_j \sim NI\left(0,\sigma_B^2\right)$	$b_j \sim NI\left(0,\sigma_B^2\right)$	σ_{CS}	$b_j \sim NI\left(0,\sigma_B^2\right)$
treatment	τ_i	τ_i	τ_i	τ_i	τ_i	τ_i	τ_i	τ_i
unit(block) | trt	σ^2		$u_{ij} \sim NI\left(0,\sigma_U^2\right)$	ϕ	$p_{ij} \sim \text{Beta}\left(\mu_{ij},\varphi\right)$	$u_{ij} \sim \Gamma\left(1/\varphi,\varphi\right)$	ϕ , φ	$u_{ij} \sim NI\left(0,\sigma_U^2\right)$ + $p_{ij} \sim \text{Beta}\left(\mu_{ij},\varphi\right)$ or $u_{ij} \sim \Gamma\left(1/\varphi,\varphi\right)$
Linear Predictor	$\mu_{ij} = \mu + \tau_i + b_j$	$\mu_{ij} = \mu + \tau_i + b_j$	$\eta_{ij} = \eta + \tau_i + b_j + u_{ij}$	$\eta_{ij} = \eta + \tau_i$	$\mu_{ij} = \mu + \tau_i + b_j$	$\mu_{ij} = \mu + \tau_i + b_j$	$\eta_{ij} = \eta + \tau_i$	$\eta_{ij} = \eta + \tau_i + b_j + u_{ij}$
distribution at unit of observation level, i.e.	$N\left(\mu_{ij},\sigma^2\right)$	Binomial$\left(N_{ij},P_{ij}\right)$ Poisson$\left(\lambda_{ij}\right)$		quasi-Binomial$\left(N_{ij},P_{ij}\right)$ quasi-Poisson$\left(\lambda_{ij}\right)$	Binomial$\left(N_{ij},P_{ij}\right)$	Poisson$\left(u_{ij}\lambda_{ij}\right)$ $\Leftrightarrow$ NB$\left(\lambda_{ij},\varphi\right)$	ill-defined quasi-likelihood	
Model Sensible or Otherwise	sensible	naive unit-level unaccounted for $\Rightarrow$ ovedispersion	sensible	sensible	sensible	sensible	nonsense redundant scale parameters at unit level	nonsense redundant distributions at unit level

adequately accounts for all sources of variation in the "how the data arise" process. Most of the time it does not and something more is needed to account for unit-level uniqueness. There are several ways to do this. Table 3 gives a summary of sensible, naive, and nonsense modeling strategies for Gaussian, binomial and discrete count data. The following expands on strategies for binomial data. Strategies for count data are analogous.

One strategy is to add a unit level random effect to the linear predictor. This is called a logit-normal model. The linear predictor is written $\eta_{ij} = \eta + \tau_i + b_j + u_{ij}$ and the associated random statement in PROC GLIMMIX is written

```
random intercept trt /
subject=block;
```

Alternatively, one can assume that the binomial probability, π_{ij} is itself a random variable. The Beta distribution is a logical choice, because π_{ij} is continuous and bounded by 0 and 1. This is called a Beta-Binomial model. The linear predictor is $\eta + \tau_i + b_j$; the Beta distribution scale parameter accounts for unit-level uniqueness, so a u_{ij} term in the linear predictor would be redundant.

Stroup (2013) discussed conditional broad inference versus marginal broad inference. The former is appropriate if inference focuses on "a typical member of the population." The latter is appropriate if inference focuses on the overall population-wide proportion. To illustrate, suppose π_{ij} is the probability of getting the flu.

Conditional broad inference means "what is the chance that I get the flu?" Marginal broad inference means "how many people are going to get the flu and what resources need to be in place to take care of them?" Conditional broad inference is typically appropriate for basic or academic research. Marginal inference is the province of policy makers.

Both the logit-normal and Beta-binomial models are examples of GLMMs that facilitate conditional broad inference. A commonly used model that accommodates marginal inference uses the linear predictor $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + \tau_i$, and a working covariance matrix

$$\begin{bmatrix} \phi+\sigma & \sigma & \sigma \\ \sigma & \phi+\sigma & \sigma \\ \sigma & \sigma & \phi+\sigma \end{bmatrix}$$ where ϕ is a scale parameter and σ is a compound symmetry working covariance. This is often called a GEE-type model. You specify this model in PROC GLIMMIX using the statements

```
model y/n = trt;
random trt / type=cs subject=block residual;
```

In this model, τ_i accounts for variation attributable to treatment, σ for block, and ϕ for unit-level uniqueness. What the logit-normal, Beta-binomial, and compound symmetry GEE models have in common is a one-to-one correspondence between WWFD sources of variation and parameters in the model. The naive model is "naive" because it lacks a parameter corresponding to unit(block). The Beta-binomial model with linear predictor $\eta + \tau_i + b_j + u_{ij}$, and Beta-binomial GEE model, are nonsense models because in both cases there is a redundant parameter associated with unit(block). This is the general strategy for using the WWFD ANOVA to identify sensible models. WWFD will not necessarily identify the best model, but it will distinguish models that make sense from those that are naive models on one hand or nonsense models on the other.

PART FOUR – BLOCKED DESIGNS II – PRECISION, POWER AND NON-GAUSSIAN DATA

Example 2. Resuming the 10 block, 3 units per block, 6 treatment example, suppose that the response variable of primary interest is binomial. In addition, suppose we know that treatments 1, 2, and 3 have success probabilities of approximately 0.10, 0.20, and 0.20 respectively. For example, treatments 1, 2 and 3 may be low, medium and high doses of a standard product. Suppose that treatments 4, 5 and 6 are low, medium and high doses of an experimental product, and conversations with the research team establish that probabilities of 0.10, 0.25 and 0.35 represent minimum plausible and clinical relevant differences from the low, medium and high dose performance of the standard. We want to know how the BIB and split-plot design compare for precision and power. From table 3, we see that the logit-normal model is a sensible model for this proposed study, and easily implemented with GLIMMIX..

To proceed, we need plausible values for the block and unit level variance. Begin with the block variance. In the logit model, it is the variance of the log of the block average odds. We can elicit this from subject matter experts by asking what are the lowest and highest plausible probabilities for treatments 1, 2 and 3, given that all blocks are not equal and these probabilities will vary from block to block. For example, treatment 1 has a probability of 0.10, but let's say that the researchers believe it could be as low as 0.05 in some blocks and as high as 0.25 in others. This means logit could be as low as $log(0.05/0.95) = -2.94$ or as high as $log(0.25/0.75) = -1.10$. Using a 6-σ approximation, the block standard deviation is approximately $(2.94 - 1.10)/6 = 0.307$ and hence the block variance is approximately 0.09. The unit variance is mathematically equivalent to the treatment×block interaction and is hence the variance of the log odds-ratio among blocks. Suppose our conversation with the research team concludes that in some blocks, there may be a negligible treatment effect but in others the odds ratio could be as high as that produced by probabilities of 0.10 and 0.35, i.e. $(0.35/0.65)/(0.1/0.9) = 4.85$. Thus, the log odds ratio could be as low as 0 or as high as 1.57. Applying the 6-σ approximation gives a standard deviation of 0.263 and a unit-level variance of approximately 0.07. The SAS statements to obtain power and precision for the BIB are

```
data bib;
input block @@;
 do eu=1 to 3; input trt @@;
   p1=0.10; p2=0.20; p3=0.20; p4=0.10; p5=0.25; p6=0.35;
   p=(trt=1)*p1+(trt=2)*p2+(trt=3)*p3+(trt=4)*p4+(trt=5)*p5+(trt=6)*p6;
   n=50; mu=n*p; output;
 end;
datalines;
 1  1 2 3
 2  1 2 4
 3  1 3 5
 4  1 4 6
 5  1 5 6
 6  2 3 6
 7  2 4 5
 8  2 5 6
 9  3 4 5
10  3 4 6
;
proc glimmix data=bib initglm;
 class block trt;
 model mu/n=trt;
 random intercept trt/subject=block;
 parms (0.09)(0.07) / hold=1,2;
 lsmestimate trt 'trt 1 v 2' 1 -1 0  0 0 0,
                 'trt 1 v 3' 1 0 -1  0 0 0,
```

```
                    'trt 4 v 5' 0 0 0  1 -1 0,
                    'trt 4 v 6' 0 0 0  1 0 -1,
                    'trt 2 v 5' 0 1 0  0 -1 0,
                    'trt 3 v 6' 0 0 1  0 0 -1 / cl exp;
  contrast 'product' trt 1 -1 0 1 -1 0,
                     trt 1 0 -1 1 0 -1;
  contrast 'dose' trt 1 1 1 -1 -1 -1;
  contrast 'product x dose' trt 1 -1 0 -1 1 0,
                            trt 1 0 -1 -1 0 1;
ods output contrasts=ftests; run;
```

The N=50 per experimental unit comes from the fact that the required number of observations to conclude a difference between probabilities 0.10 and 0.20 to be significantly different at the 0.05 level with power 0.8 is 250, so an obvious question is, if we replicate each treatment in 5 blocks, is it okay to split the 250 observations on each treatment 50 per block? MU=P*N is the response variable, i.e. the number of expected successes in a given experimental unit. The LSMESTIMATE statement defines selected pairwise differences corresponding to simple effects of likely interest. The EXP option converts differences, which are on the model scale, to odds-ratios, and obtains their expected confidence interval width. The ODS OUTPUT statement creates a data set with the expected *F*-value. As noted in Example 1, multiplying it by the numerator degrees of freedom gives the non-centrality parameter. Use the following statement to obtain power.

```
data power;
 set ftests;
  alpha=0.05;
  phi=numdf*fvalue;
  fcrit=finv(1-alpha,numdf,dendf,0);
  power=1-probf(fcrit,numdf,dendf,phi);
proc print data=power;
run;
```

You can modify the programs to obtain precision and power for the other designs under consideration. If you include the complete block design in the comparison, you must determine the change in block and unit-level variance that results from using a different block size. As with the Gaussian example, you can use simulation to do this. Here, the conversion yields $\sigma^2_B = 0.05$ $\sigma^2_U = 0.12$ as the parameters to use if you evaluate a complete block design. Output 5 shows the precision and power results for the BIB and SPLT (disconnected/group-nested/split-plot) designs

Output 5. Precision (standard error and confidence limits of odds-ratios for selected pairwise comparisons) and Power for factorial effects. BIB versus Split-Plot design.

Obs	Label	stderr_bib	stderr_splt	LCL_bib	UCL_bib	LCL_splt	UCL_splt
1	trt 1 v 2	0.32805	0.31216	0.22088	0.89430	0.22931	0.86142
2	trt 1 v 3	0.32856	0.31216	0.22063	0.89529	0.22931	0.86142
3	trt 4 v 5	0.32299	0.30623	0.16745	0.66354	0.17416	0.63799
4	trt 4 v 6	0.31695	0.30004	0.10501	0.40550	0.10924	0.38980
5	trt 2 v 5	0.29185	0.33216	0.40263	1.39707	0.37089	1.51661
6	trt 3 v 6	0.28526	0.32647	0.25277	0.85279	0.23239	0.92759

Obs	Label	power_bib	power_splt
1	product	0.99418	0.99744
2	dose	0.44466	0.25403
3	product x dose	0.27484	0.30812

In this case, the BIB is clearly the better choice. The split-plot has only slightly better precision than the BIB for comparisons within the whole-plot (1 vs 2, 1 vs 3, 4 vs 5, 4 vs 6) and power for the product × dose interaction, but is more noticeably inferior in all other respects. Keep in mind that this is *not* a blanket endorsement of the BIB versus the split-plot for binomial data. Different parameters will yield different results. The **two most important lessons** are 1) do not use the results of a precision and power analysis for a Gaussian response variable to make decisions about

a design when the primary response variable of interest is not Gaussian, and 2) 250 observations per treatment may be adequate for a simple design, but when the observations are spread among multiple units, and those units contribute additional variability, power drops dramatically. To achieve 80% power for the test of interaction, for example, you would need to replicate the split-plot design several times and use considerably more than 250 observations per treatment.

Example 3: What about count data?

You can modify the binomial program to plan when the primary response variable is a discrete count. We suggest using the negative binomial and not the Poisson. In general, negative binomial counts arise from a process more consistent with the way most count data arise. Studies from a number of disciplines consistently show that the negative binomial provides a better fit. This is especially true of biological count data. Finally, power analysis using the Poisson is much more prone to egregiously optimistic recommendations because Poisson GLMMs are especially prone to overdispersion.

To illustrate precision and power analysis for count data, suppose in the last example the primary response variable was a discrete count instead of a binomial proportion. Suppose the treatments have the same designations: treatment group {1,2,3} and {4,5,6} represent the standard product and experimental product, respectively, and the levels within each product are low, medium and high dose. A discussion with the research team establishes expected mean counts for the standard product as 10, 10, and 20 for treatments 1, 2 and 3, respectively, and 5, 30 and 30 as expected means across doses corresponding to the minimum clinically relevant difference for the experimental product versus the standard. This sets the means to be defined in the exemplary data set.

For the GLIMMIX step, recalling that the linear predictor for the negative binomial with a blocked design is $\log(\lambda_{ij}) = \eta + \tau_i + b_j$ where b_j denotes the random block effect, you need to specify the block variance and the negative binomial scale parameter. The block variance measures variability in the log mean counts from block to block. For example, consider blocks that receive all doses of the standard product. The block mean is $(1/3)(10+10+20) \cong 13.3$. Suppose that the research team believes that there could be blocks with a mean as low as 5 and as high as 35. On the link scale, the range is $\log(5) = 1.61$ to $\log(35) = 3.56$. Using the 6-σ approximation gives a block variance of $\left((3.56-1.61)/6\right)^2 \cong 0.11$. For the scale parameter, begin with the negative binomial variance, $\lambda + \varphi\lambda^2$. Suppose that in addition to knowing the mean for the low dose of the standard product, researchers believe that an observed count of greater than 30 would be rare. For the high dose, they know the mean is approximately 20 and anything above 50 would be rare. With these characteristics in mind, you can evaluate the negative binomial c.d.f. directly, or obtain an empirical distribution by simulation. Either way, a scale parameter of φ approximately equal to 0.4 yields a negative binomial distribution with these characteristics. The resulting exemplary data and GLIMMIX steps are

```
 data bib;
input block @@;
 do eu=1 to 3;
  input trt @@;
   mu1=10; mu2=10; mu3=20; mu4=5; mu5=30; mu6=30;
   mu=(trt=1)*mu1+(trt=2)*mu2+(trt=3)*mu3+(trt=4)*mu4+(trt=5)*mu5+(trt=6)*mu6;
   output;
 end;
datalines;
 1   1 2 3
 2   1 2 4
 3   1 3 5
 4   1 4 6
 5   1 5 6
 6   2 3 6
 7   2 4 5
 8   2 5 6
 9   3 4 5
10   3 4 6
;
proc glimmix data=bib initglm;
 class block trt;
 model mu=trt/d=negbin;
 random intercept /subject=block;
 parms (0.11)(0.4) / hold=1,2;
```

```
lsmestimate trt 'trt 1 v 2' 1 -1 0  0 0 0,
                'trt 1 v 3' 1 0 -1  0 0 0,
                'trt 4 v 5' 0 0 0  1 -1 0,
                'trt 4 v 6' 0 0 0  1 0 -1,
                'trt 2 v 5' 0 1 0  0 -1 0,
                'trt 3 v 6' 0 0 1  0 0 -1 / cl exp;
contrast 'product' trt 1 -1 0 1 -1 0,
                   trt 1 0 -1 1 0 -1;
contrast 'dose' trt 1 1 1 -1 -1 -1;
contrast 'product x dose' trt 1 -1 0 -1 1 0,
                          trt 1 0 -1 -1 0 1;
ods output contrasts=ftests lsmestimates=bib_diff;
```

These statements are similar to the statements for binomial data. The only differences are 1) the response variable, (expected count, a.k.a. λ) is called MU, and 2) the distribution, D=NEGBIN, must be specified in the MODEL statement. Also, because the link function is the log, applying the EXP option to differences estimated on the model scale yields the ratio $\hat{\lambda}_i/\hat{\lambda}_{i'}$ for the i^{th} and i'^{th} treatments. If you choose to include a complete block design in the comparison, run a simulation to obtain converted variance components. In this example, the converted parameters are $\sigma_B^2 = 0.07$ and $\varphi = 0.41$. Results for the BIB and split-plot (SPLT) designs appear in Output 6.

Output 6. Precision (standard error and confidence limits of mean ratios for selected pairwise comparisons) and Power for factorial effects. BIB versus Split-Plot design.

Obs	Label	stderr_bib	stderr_splt	LCL_bib	UCL_bib	LCL_splt	UCL_splt
1	trt 1 v 2	0.46565	0.44721	0.37064	2.69801	0.38750	2.58067
2	trt 1 v 3	0.45511	0.43589	0.18953	1.31903	0.19846	1.25973
3	trt 4 v 5	0.47288	0.45461	0.06083	0.45665	0.06358	0.43691
4	trt 4 v 6	0.47288	0.45461	0.06083	0.45665	0.06358	0.43691
5	trt 2 v 5	0.45134	0.48028	0.12738	0.87231	0.12042	0.92268
6	trt 3 v 6	0.44012	0.46975	0.26092	1.70340	0.24628	1.80465

Obs	Label	power_bib	power_splt
1	product	0.88595	0.91429
2	dose	0.15899	0.12042
3	product x dose	0.58411	0.62605

Similar lessons apply. The biggest take home message is that all of the precision and power analyses that appear in this paper *depend absolutely on using GLMM theory, methods and software*. You have to be able to properly account for the role of block and unit-level variation, the presence of recovery of inter-block information in some designs but not others, and the impact of changing block size and block configuration on the relative magnitude of block and unit-level variance. Without GLMM thinking, all of this is missed, and egregious miscalculations of design and sample size requirement can be and often are the result.

PART FIVE – WWFD, COMPLEX DESIGNS AND MODEL CONSTRUCTION

Up to this point, we have only considered designs with one blocking criterion and one size experimental unit. For these designs, the WWFD process and model construction are relatively easy. However, many study designs have much more complex structures. In fact, it is often a challenge to determine exactly what design structure was actually used. This is where techniques for accurate model construction become critical, and the opportunity increases dramatically for modeling issues to occur that are really the result of misdiagnosed designs. It would be impossible to cover every contingency with which data analysts might be faced. In this section, we work through an example intended to illustrate many of the complications, and the strategies for dealing with them, that readers can expect to encounter.

Example 4: Figure 3 shows a design conducted to evaluate the effects of two treatment factors, referred to as factor A and factor B. Four “reps” running the length of the experimental space are each divided into 2 strips. One level of factor A is assigned to each strip. Each “rep” is divided into 3 “blocks.” Each “block” is divided into two “plots.” Within

each block, one level of factor B is randomly assigned to each plot. The treatment design is a 2×2 factorial. The "topographical" or "experiment" design is more difficult to characterize: it has the signature characteristics of a strip-split-plot design, but it is much more complex and has twists not commonly seen in textbook strip-split-plot designs.

Figure 3. Complex 2×2 factorial Strip-Split Plot Study Design

Rep 1	Block 1		Block 2		Block 3	
	B_1	B_2	B_1	B_2	B_1	B_2
Rep 2	Block 4		Block 5		Block 6	
	B_2	B_1	B_1	B_2	B_2	B_1
Rep 3	Block 7		Block 8		Block 9	
	B_2	B_1	B_2	B_1	B_1	B_2
Rep 4	Block 10		Block 11		Block 12	
	B_2	B_1	B_2	B_1	B_1	B_2

The strips across each "half-rep" show where the levels of factor A are assigned. The strip shaded gray shows where A_1, the first level of A, has been assigned. The unshaded strips show where A_2 has been assigned. The WWFD process requires two "topographical" or "experiment design" steps. First, consider only the "reps" and the strips to which the levels of factor A are assigned. Figure 4 shows this aspect of the design.

Figure 4. "Rep" and strip within Rep aspect of complex design

Rep 1	
Rep 2	
Rep 3	
Rep 4	

The ANOVA sources of variation for this aspect can be written

Source	**d.f.**
rep	3
strip(rep)	4
Total for rep + strip aspect	**7**

Now consider the block and plot within block aspect. Visualize this aspect using Figure 5.

Figure 5. Block within Rep and Plot within Block Aspect

Rep 1	Block 1 B_1 | B_2	Block 2 B_1 | B_2	Block 3 B_1 | B_2
Rep 2	Block 4 B_2 | B_1	Block 5 B_1 | B_2	Block 6 B_2 | B_1
Rep 3	Block 7 B_2 | B_1	Block 8 B_2 | B_1	Block 9 B_1 | B_2
Rep 4	Block 10 B_2 | B_1	Block 11 B_2 | B_1	Block 12 B_1 | B_2

The ANOVA sources of variation for this aspect can be written

Source	**d.f.**
rep	3
block(rep)	8
plot(block, rep)	12
Total for rep-block-plot aspect	**23**

Now, combine the two "experiment design" ANOVAs.

Source	**d.f.**
rep	3
strip(rep)	4
block(rep)	8
strip×block(rep)	8
plot(block, rep)	12
strip×plot(block, rep)	12
Total	**47**

Finally, augment this table with the treatment design, being careful to place treatment factors in the row above the unit to which they were applied (e.g. A to strip, B to plot). Integrate to form the combined table.

experiment design		**treatment design**		**combined**	
Source	**d.f.**	**Source**	**d.f.**	**Source**	**d.f.**
rep	3			rep	3
		A	1	A	1
strip(rep)	4			strip(rep) = rep*A	4-1=3
block(rep)	8			block(rep)	8
strip×block(rep)	8			strip×block(rep) = A*block(rep)	8
		B	1	B	1
plot(block, rep	12			plot(block, rep) = B*block(rep)	12-1=11
		A×B	1	A×B	1
strip×plot(block, rep)	12	"parallels"	44	strip×plot(block, rep) | A,B	12-1=11
Total	**47**		**47**	**Total**	**47**

When this table was completed, there was a lively discussion among statisticians with considerable consulting experience about whether "strip×block(rep)" should be a distinct term in the ANOVA. At the end of the day, we concluded that it was indeed a distinct physical entity in the design and, when in doubt, respect the design.

Following Table 3, the next steps are

- Write the distribution of the observations at the unit level – i.e. the strip×plot(block, rep) level.
- Decide what terms in the rows above the unit level have probability distributions. Presumably, REPs are a sample of a larger population. If so, all terms originating in the experiment design column will be modeled by random effects, and all effects originating in the treatment design column will be modeled as fixed.
- If the distribution is not Gaussian, make a decision along the lines of the discussion accompanying Table 3 as to how you will account for variation attributable to uniqueness at the unit of observation level.
- Finally, decide what link function and linear predictor you will use.
- If there are repeated measures, either in space or time, the observations at the unit level may be a vector of correlated observations with a covariance structure. This may be part of the distribution of the observations (in GLIMMIX language, the ***R*** matrix), e.g. for Gaussian data, or part of the linear predictor (in GLIMMIX language, ***G***-side covariance), e.g. for binomial or Poisson data.

As an example, if the data are discrete counts, sensible models could include the Poisson-normal or negative binomial model, assuming that target of inference is λ_{ij} , the A×B mean for a typical member of the population . If the marginal mean is the target, a hybrid, i.e. a model that is part GLMM, part quasi-likelihood might be chosen. The GLIMMIX statements for the negative binomial model are

```
proc glimmix;
 class rep block a b;
 model count = a|b / d=negbin;
 random intercept a block a*block b*block / subject=rep;
```

For a marginal model, a plausible program might be

```
proc glimmix;
 class rep block a b;
 model count = a|b / d=poisson;
 random intercept a block a*block / subject=rep;
 random a / type=cs subject=rep*block*b residual;
```

Mastering the WWFD process, and understanding the probability or quasi-likelihood structures that make sense in the context of the resulting ANOVA table, are essential for determining a model consistent with the study design.

CONCLUSION

This presentation focused on two variations of the design-modeling interface. One concerns extracting sensible models that are consistent with design structure, as broadly defined. The other concerns the use of sensible models to plan future studies.

Competent statistical modeling requires a firm command of design concepts. Disconnect between design and model is a prolific breeding ground for modeling problems. The WWFD process, which is, in essence, the ANOVA table re-purposed for the 21st century in the service of generalized and mixed model construction, is a useful tool to clarify the nuances one needs to consider when translating *how the data were collected* to *how the data will be modeled*. Traditional modeling curriculum often misses these nuances. WWFD is a useful teaching tool and a useful aid to consulting, In addition, WWFD can be used to frame research questions about GLMMs when they are used in uncharted territory.

Conventional power and sample size methodology is firmly rooted in ordinary least squares, with little or no recognition of the impact of random model effects, and even less recognition of the combined impact of random model effects *in conjunction with* non-Gaussian data on precision and power assessment. Among those who do recognize what is involved, there is widespread – and often erroneous – belief that elaborate methods, often simulation-based, are necessary. The methods presented in this section yield precision and power results that are easy to compute, once you understand the modeling and design principles involved. For Gaussian, binomial and count data, these methods are demonstrably accurate and easily supported by corroborating simulation results. The take home message of the examples presented in this paper is that power and precision assessment of proposed designs with any degree of complexity – which essentially means any form of blocking – cannot be done accurately without the use of GLMM thinking and GLMM methodology.

As a final note, the author fervently urges those who teach modeling, design, and statistical methods courses to incorporate these ideas into their classroom material.

REFERENCES

1. Federer, W.T. (1955) *Experimental Design – Theory and Applications*. New York: McMillan Publishing.
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. Lohr, V.I. and O'Brien, R.G., (1984) Power analysis for univariate linear models: SAS Makes It Easy. SAS Users' Group International: Proceedings of the Ninth Annual Conference. Cary, NC: SAS Institute, Inc. pp. 847-852.
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. Stroup, W.W. (1999) Mixed Model Procedures to Assess Power, Precision and Sample Size in Design of Experiments. 1999 Proceedings of the Biopharmaceutical Section. Alexandria, VA. American Statistical Association. pp. 15-24.
8. Stroup, W.W. (2011) Living with Generalized Linear Mixed Models. *SAS Global Forum 2011.* Paper 349-2011.
9. Stroup, W.W. (2013) *Generalized Linear Mixed Models: Model Concepts, Methods and Applications.* Boca Raton, FL: CRC Press.
10. SAS Institute, Inc. (2011) *SAS/STAT® 9.4 User's Guide.* Cary, NC: SAS Institute, Inc.
11. Yates, F. (1935) Complex Experiments (with discussion), *Supplement to the Journal of the Royal Statistical Society* **2** pp. 181-247.

CONTACT INFORMATION

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