Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

2013 - 25th Annual Conference Proceedings

NON-NORMAL DATA IN AGRICULTURAL EXPERIMENTS

W. W. Stroup

NON-NORMAL DATA IN AGRICULTURAL EXPERIMENTS

W.W. Stroup, University of Nebraska, Lincoln

ABSTRACT

Advances in computers and modeling over the past couple of decades have greatly expanded options for analyzing non-normal data. Prior to the 1990's, options were largely limited to analysis of variance (ANOVA), either on untransformed data or after applying a variance stabilizing transformation. With or without transformations, this approach depends heavily on the Central Limit Theorem and ANOVA's robustness. The availability of software such as R's *lme4* package and SAS® PROC GLIMMIX changed the conversation with regard to non-normal data. With expanded options come dilemmas. We have software choices – R and SAS among many others. Models have conditional and marginal formulations. There are GLMMs, GEEs among a host of other acronyms. There are different estimation methods – linearization (e.g. pseudo-likelihood), integral approximation (e.g. quadrature) and Bayesian methods. How do we decide what to use? How much, if any, advantage is there to using GLMMs or GEEs versus more traditional ANOVA-based methods? Stroup (2013) introduced a design-to-model thought exercise called WWFD (What Would Fisher Do). This paper illustrates the use ofWWFD to clarify thinking about plausible probability processes giving rise to data in designed experiments, modeling options for analyzing non-normal data, and how to use the two evaluate small-sample behavior of competing options. Examples with binomial and count data are given. While the examples are not exhaustive, they raise issues and call into question common practice and conventional wisdom regarding non-normal data in agricultural research.

1. INTRODUCTION

Once, before there was a Conference on Applied Statistics in Agriculture, analyzing non-normal data from designed experiments seemed to be a settled issue. For most of the past century "standard statistical methods" in agricultural research equated to analysis of variance (ANOVA).Because ANOVA focuses on means, statistical analysis placed heavy reliance on the Central Limit Theorem: given sufficient experimental units per treatment – read "properly designed experiment" – sample meanscould be assumed to have an approximate normal distribution.

A major problem with this line of reasoning stemmed from potential heterogeneity of variance. For example, with binomial data, count data, and other non-normal data common in agricultural research, the variance is a function of the mean; if the mean differs by treatment, so must the variance. Bartlett (1947) proposed variance-stabilizing transformations for distributionscommonly encountered in experimental research. Use of transformationsbecame standard operating procedure for many agricultural disciplines. Most statistical methods texts, e.g. Snedecor and Cochran (1989) or Steel and Torrie (1980), include a section on transformations Bartlett proposed, often with minor variations.

With or without transformations, reliance on ANOVA seemed further justified by its robustness. In his book *Beyond ANOVA*, Miller (1997 – originally published 1986) provided an in depth exploration of ANOVA assumptions and the extent to which they had to be violatedto render ANOVA results unreliable or not credible. Reading Miller and related works seemed to vindicate the research community's confidence in ANOVA.

Things became more complicated in the 1990s, when theory and methods that had been incubating for decades reached a tipping point of awareness, availability and practicality.Early in the 1970s, Nelder and Wedderburn (1972) introduced generalized linear models (GLMs). GLMs required only assuming that the data distribution belongs to the exponential family, or, even less restrictively, to the family of quasi-likelihoods. ANOVA with fixed effectsis a special case of the generalized linear model that assumes a normal distribution with homoscedastic variance. Linear mixed model (LMM) methodology appeared even earlier. Yates (1940) introduced recovery of inter-block information, a crucial precursor of LMM analysis. Eisenhart (1947) introduced formal conceptual distinctions between fixed-effects-only and mixed-effects models. Henderson (1953, 1963), and subsequently Harville (1976, 1978) did seminal work essential to the development of modern LMM theory and methods.Prior to the 1980s, LMMs were seen as the concern of only a few highly specialized applications, notably plant and animal genetics. Laird and Ware (1982) and a USDA-supported regional project, S-189 (1989) brought mixed model methods to the attention of larger research communities, including agriculture. This awareness included the realization that ANOVA for multi-level designs, e.g. split-plots, is best implemented using LMM methods.Before the last decade of the 20th Century, GLMs and LMMs were not practical because of limitations in computer capability and the absence of useable software. This changed in the early 1990s with the introduction of software such as SAS® PROC MIXED and PROC GENMOD. Thanks in part to the efforts of Applied Statistics in Agriculture conference participants, by the mid-1990s LMMs had become mainstream in agricultural circles and GLMs were well-known, if not quite mainstream.

The appearance of GLM and LMM software created a dilemma with regard to non-normal data. With designs for which fixed-effects-only models were appropriate, data analysts now had a choice among ANOVA without transformation, ANOVA with transformation, or GLM-based analysis. Analyses from these three methods could – and often did – producemutually exclusive results. Which result should the researcher report and how was the choice to be justified to referees and reviewers? For designs that called for LMM-based analysis – e.g. split-plot or repeated measures experiments, or data with spatial correlation – non-normal data raised another issue. How would transformations – if they seemed warranted – interact with random model effects? If random effects were approximately normal on the original data scale, would transforming the data distort the random effect distribution and compromise the analysis?

Things became even more complicated in the 2000s. Generalized linear mixed model (GLMM) theory advanced rapidly in the 1990s.These developments and the continued improvement in computers allowed, in the mid-2000s, the introduction of GLMM software such as R's *lme4* package and SAS PROC GLIMMIX.Today's data analyst is faced with a wide variety of options for analyzing non-normal data from designs that call for a mixed-model approach. Should one use a GLMM, an LMM on transformed data, an LMM on untransformed data, or something else? What is needed is a way to think about the problem, a way to make sensible and defensible choices among competing options.

The purpose of this paper is to present an approach that facilitates making comparisons among GLMM- and ANOVA-based methods and to illustrate ways in which this approach might be employed. The technique demonstrated here can be used both as a research tool to compare different methods of analysisand as a teaching tools – e.g. to help linear model students learn how to think about modeling complex experiments or to help consulting clients understand the many ways that sources ofvariation affect can affect the data they observe. The illustrations are meant to be examples. They are not intended to be exhaustive and should not be interpreted as a comprehensive comparison of GLMM vs. ANOVA vs. transformations, although the cases considered here certainly speak to the GLMM vs. ANOVA vs. transformation issue and contain ample food for thought.

Although the examples in this paper are limited to three cases of interest, the *approach* shown in this paper *could* be implemented in an exhaustive manner to provide the basis for a comprehensiveGLMM vs. ANOVA vs. transformation vs. other-alternatives-of-interest comparison. The author invites interested readers to do so.

2. THE APPROACH: DEFINING THE MODELING PROBLEM

Littell, et al. (2006) describe statistical models as "...mathematical descriptions of how data conceivably can be produced." Beginning students and unsophisticated practitioners of statistics tend to regard the model exclusively as a template for data analysis. In doing so they miss a crucial – arguably the most crucial – aspect of modeling: "how data conceivably can be produced."That is, in terms of sources of variation and probability distributions,what are plausible narratives – typically there is more than one – that might explain how the observed data arose? Addressing *both* "how did the data arise?" *and* "how will the data be analyzed?" are essential. Aligning the two is the key to informative statistical modeling. The data creation narrative is especially important for designing a simulation study to provide a meaningful comparison among competing methods of analysis.

2A. What Would Fisher Do – A Review

Stroup (2013) presented a gimmick entitled WWFD (What Would Fisher Do?) as a technique to help linear models students and statistical consultants translate the description of a study design into a plausible statistical model. WWFD is based on comments by Fisher following a presentation by Yates (1935) to the Royal Statistical Society. Fisher said that any study could be characterized in terms of its "topographical" and "treatment" aspects. Federer (1955) would later refer to these as the "experiment design" and the "treatment design," respectively. Fisher said that if one wrote separate ANOVA sources of variation and degrees of freedom for each aspect and then combined them, it should be clear how to proceed with the analysis. WWFD starts with this ANOVA-based process and adds a GLMM twist. The approach is similar to strategy employed for messy data by Milliken and Johnson (2008).

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, but do not necessarily assume a complete block design.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**	$bu-1$

The placement of the treatment source of variation matters. Notice that it is 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 everything else that treatment sources of variation did not account for in the “treatment” ANOVA. As will be clear below, “parallels” play no role in the combined ANOVA.

The combined ANOVA appears as follows.

“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 read the combined ANOVA in model form as $y_{ij} = \mu + b_i + \tau_j + e_{ij}$, where there is a one-to-one correspondence between subscripted terms on the right-hand side and terms listed under “Source of Variation.” That is, block $\Rightarrow b_i$, treatment $\Rightarrow \tau_j$ and e.u.(block) |trt (read “unit within block after accounting for treatment”) a.k.a. “residual” $\Rightarrow e_{ij}$. However, GLMM textbooks call this the “model equation” approach, and point out that it obstructs constructing a plausible model if the data are not Gaussian. Instead, 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 thesource of variation that appears 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, $y_{ij} \sim NI\left(\mu_{ij}, \sigma^2\right)$. On the other hand, if the observations are binomial with N independent binary observations per experimental unit, i.e. the observations are 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 i^{th} block, j^{th} treatment.

The next model-construction step involves deciding how the other sources of variation affect the subscripted parameter of the unit-level distribution. In other words, one needs to choose a link function and a linear predictor. For example, for Gaussian data, the standard link is the identity and the standard linear predictor for a randomized block design is $\mu_{ij} = \mu + b_i + \tau_j$; for binomial data, the standard link is the logit, $\text{logit}\left(\pi_{ij}\right) = \log\left[\pi_{ij} \big/ \left(1-\pi_{ij}\right)\right] = \eta_{ij} = \eta + b_i + \tau_j$.

The final model-construction 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 back to that population, then, by definition, one has a strong case for associating a probability distribution with the block effects.

In principle, WWFD can be applied to designs of arbitrary complexity. For example, a split-plot design with whole-plot units applied in randomized blocks – *r* blocks, *w* whole plot units per block, *s* split plot units per whole plot, *a* levels of the whole plot factor (factor A), and *b* levels of the split plot factor (factor B) – would have the following WWFD ANOVA:

Topographical		**Treatment**		**Combined**	
Source	**d.f.**	**Source**	**d.f.**	**Source**	**d.f.**
block	$r-1$			block	$r-1$
		A	$a-1$	A	$a-1$
whole plot unit (block)	$r\left(w-1\right)$			w.p.u(blk) | A	$r\left(w-1\right)-\left(a-1\right)=$ $rw-a$
		B	$b-1$	B	$b-1$
		A × B	$\left(a-1\right)\left(b-1\right)$	A × B	$\left(a-1\right)\left(b-1\right)$
split plot unit (w.p.u.)	$rw\left(s-1\right)$	“parallels”		s.p.u.(w.p.) | B,A a.k.a. “residual”	$rw\left(s-1\right)-a\left(b-1\right)$
TOTAL	$rws-1$	**TOTAL**	$rws-1$	**TOTAL**	$rws-1$

After writing the ANOVA, one would need to assign a plausible distribution to observations at the unit level – e.g. $y_{ijk} \mid r_i, w_{ij} \sim NI\left(\mu_{ijk}, \sigma^2\right)$ for independent Gaussian dataor $y_{ijk} \mid r_i, w_{ij} \sim \text{Binomial}\left(N, \pi_{ijk}\right)$ for binomial data – then a link and a linear predictor, e.g. $\eta_{ijk} = \eta + r_i + \alpha_j + w_{ij} + \beta_k + \left(\alpha\beta\right)_{jk}$, where block $\Rightarrow r_i$, A $\Rightarrow \alpha_j$, whole plot $\Rightarrow w_{ij}$, B $\Rightarrow \beta_k$ and $\text{A} \times \text{B} \Rightarrow \left(\alpha\beta\right)_{jk}$. Obviously, at least the whole plot unit effect, w_{ij}, would have to be defined as a random effect with an assumed probability distribution.

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. Typically in GLMM theory **y** denotes the observation vector, **b** denotes the random effect vector, and $f\left(\mathbf{y} \mid \mathbf{b}\right)$ denotes the p.d.f. of the unit-level distribution.
- write the link function $\boldsymbol{\eta} = g\left(\boldsymbol{\mu} \mid \mathbf{b}\right)$, where $\boldsymbol{\mu} \mid \mathbf{b} = E\left(\mathbf{y} \mid \mathbf{b}\right)$
- 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 $f\left(\mathbf{b}\right)$.

2B. Using WWFD to Describe How Data Arose

The obvious use of the WWFD device is to define a GLMM to be used for data analysis. However, as we will see in Section 2C, WWFD can lead to several models within the GLMM family, each of which plausibly follows from the model-construction process. In order to compare competing models, we need to describe a process by which we think the data arose and then see how well each model estimates or tests aspects deemed most important. In this section, we show how to use WWFD to conceptualize how the data arose. We use the randomized block to illustrate.

The combined WWFD ANOVA for the randomized block design is

Combined		
Source of Variation	**d.f.**	**Implied p.d.f.**
block	$b-1$	$f\left(\mathbf{b}\right)$
treatment	$t-1$	none if trt effect fixed
e.u. (block) | trt a.k.a. block × trt a.k.a. “residual”	$bu-b-t+1$	$f\left(\mathbf{y}\mid\mathbf{b};\boldsymbol{\mu}\mid\mathbf{b}\right)$
TOTAL	$bu-1$	

The process that follows can be described in stages. First, the mean of the unit-level distribution, $\boldsymbol{\mu}\mid\mathbf{b}$ must depend in some way on the block and treatment effects. Once the unit-level mean is determined, the observations arise from a distribution with p.d.f. $f\left(\mathbf{y}\mid\mathbf{b};\boldsymbol{\mu}\mid\mathbf{b}\right)$.

If the observations are assumed to be Gaussian, this is straightforward: $\mu_{ij}=\mu+b_i+\tau_j$. Assuming the block effects are random, they follow a probability distribution, e.g. $b_i \sim NI\left(0,\sigma_B^2\right)$. The observations then arise according to a $NI\left(\mu+b_i+\tau_j,\sigma^2\right)$ distribution.

If the observations are assumed to be non-Gaussian, the process leading to the unit-level mean can take several forms. Two obvious possibilities are

- $\mu_{ij}\mid b_i=g^{-1}\left(\eta+b_i+\tau_j\right)$, where $g^{-1}\left(\bullet\right)$ denotes the inverse link function
- $\mu_{ij}\mid b_i=\eta+b_i+\tau_j$

In the former, block and treatment effects perturb the unit-level mean on the scale defined by the link function. For example, this would describe a scenario in which the block and treatment directly affect the canonical parameter of the distribution. In the latter, the additive effects of block and treatment perturb the unit-level mean directly. In some cases with direct additive effects, one would have to place boundaries on $\eta+b_i+\tau_j$ to keep μ_{ij} within the parameter space. For example, for binomial data, $\eta+b_i+\tau_j$ is not constrained to be between 0 and 1 but the unit-level binomial mean, commonly denoted $\pi_{ij} \mid b_i$, obviously cannot be <0 or >1. Variations on this theme could include

- $\mu_{ij} \mid b_i = g^{-1}\left(\eta + b_i\right) + \tau_j$
- $\mu_{ij} \mid b_i = g^{-1}\left(\eta + \tau_j\right) + b_i$

While one or more of the above might be more plausible in a given context, there is no reason in theory to preclude any of the above. This is important, because if one wants to do a comprehensive comparison among competing models, their behavior under each of the above scenarios should be evaluated.

Once one defines the way in which blocks and treatments perturb the unit-level mean, then one can trace the process by which observations arise. First, block effects arise via $f\left(b_i\right)$. Note that under standard GLMM theory $f\left(b_i\right)$ is Gaussian, but there is no reason why block effects must arise according to a Gaussian distribution. In fact, it is of great interest in understanding the small-sample behavior of GLMMs to ask what happens when block effects do *not* have a Gaussian distribution. Once the block effects arise, they, along with the τ_j determine the unit level mean $\mu_{ij} \mid b_i$ and the observations subsequently arise according to $f\left(y_{ij} \mid b_i; \mu_{ij} \mid b_i\right)$.

2C. Using WWFD to Define Modeling Options

By "modeling options" we mean different GLMMs one might use for data analysis. To illustrate, consider a randomized block design with binomial data. Three obvious modeling options are

- **Option 1:** assume the sample proportion $p_{ij} = y_{ij} / N$ has an approximate normal distribution and fit the model $p_{ij} = \mu + b_i + \tau_j + e_{ij}$, where $e_{ij} \sim NI\left(0, \sigma^2\right)$
- **Option 2:** use the arc sine square root transformation, i.e. $\sin^{-1}\left(\sqrt{p_{ij}}\right) = \mu + b_i + \tau_j + e_{ij}$ where $e_{ij} \sim NI\left(0, \sigma^2\right)$
- **Option 3:** basic GLMM, i.e. $y_{ij} \mid b_i \sim \text{Binomial}\left(N, \pi_{ij}\right)$; link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + b_i + \tau_j$, $b_i \sim NI\left(0, \sigma_B^2\right)$

Properly understood, the WWFD ANOVA reveals a potential problem with Option 3, the basic GLMM. Specifically, unlike the normal distribution, the binomial only has one parameter. With a Gaussian linear model, knowing the model parameters yields $\mu + b_i + \tau_j = \mu_{ij}$, but μ_{ij} says nothing about the unit-level variance. On the other hand, in the binomial case, knowing the model parameters yields $1 \Big/ \left\{1 + \exp\left[-\left(\eta + b_i + \tau_j\right)\right]\right\} = \pi_{ij}$, which also determines the unit level variance, $\pi_{ij}\left(1 - \pi_{ij}\right) / N$. In the Gaussian case, the "residual" line of the ANOVA contains the information needed to estimate σ^2 and also accounts for any experimental unit uniqueness not attributable to the block and treatment effects. In the basic binomial GLMM, Option 3, the last line is not used. If non-negligible unit-level uniqueness exists, the model contains nothing to account for it. Option 3 is prone to overdispersion as a result of this oversight.

Several options exist to account for unit-level uniqueness in the binomial GLMM.

- **Option 4:** add a block × treatment term to the linear predictor. The amended GLMM is $y_{ij} \mid b_i, (bt)_{ij} \sim \text{Binomial}\left(N, \pi_{ij}\right)$; link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + b_i + \tau_j + (bt)_{ij}$, $b_i \sim NI\left(0, \sigma_B^2\right)$, $(bt)_{ij} \sim NI\left(0, \sigma_{BT}^2\right)$. The block × treatment term is interpreted as unit-level uniqueness.
- **Option 5:** reparameterize Option 4 as a compound symmetry model: $y_{ij} \mid (bt)_{ij} \sim \text{Binomial}\left(N, \pi_{ij}\right)$; link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + \tau_j + (bt)_{ij}$, $\begin{pmatrix} (bt)_{i1} \\ (bt)_{i2} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_{CS}^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$. This would be especially useful if Option 4 yielded negative variance component estimates.
- **Option 6:** Add a unit-level scale parameter to the basic GLMM in Option 3. That is, $y_{ij} \mid b_i \sim quasi - \text{Binomial}\left(N, \pi_{ij}\right)$; link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + b_i + \tau_j$, $b_i \sim NI\left(0, \sigma_B^2\right)$, $quasi - \text{Var}\left(y_{ij} \mid b_i\right) \propto \phi \pi_{ij}\left(1 - \pi_{ij}\right)$
- **Option 7:**Replace the compound symmetry covariance structure in Option 5 with a working covariance. That is, $y_{ij} \sim quasi - \text{Binomial}\left(N, \pi_j\right)$; link $\eta_j = \text{logit}\left(\pi_j\right) = \eta + \tau_j$, $quasi - \text{Var}\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \propto \pi_{ij}\left(1 - \pi_{ij}\right)\phi \begin{bmatrix} 1 & \rho_w \\ \rho_w & 1 \end{bmatrix}$ where ρ_w denotes working correlation. This model targets the mean of the marginal distribution of y_{ij}, not π_j, and is variously called a *marginal* GLMM (whereas Options 4 and 5 are called *conditional* GLMMs) or a GEE model.
- **Option 8:** Similar to Option 4 except assume that π_{ij} is perturbed at the unit-level by a non-Gaussian random effect. The beta-binomial is the most important example of such a model. Assume at the unit-level, $\pi_{ij} \mid b_i \sim Beta\left(c_{ij}, d_{ij}\right)$. Notice that this means the parameters of the beta distribution depend on block and treatment. Assuming constant*N*, one can fit this model with GLMM software as follows. Use the sample proportion $p_{ij} = y_{ij} / N$. Assume $p_{ij} \mid b_i \sim Beta\left(\pi_{ij}, \varphi\right)$, using the GLMM-friendly reparameterization of the Beta from Ferrari and Cribari-Neto (2004). The link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + b_i + \tau_j$, where $b_i \sim NI\left(0, \sigma_B^2\right)$

Other options undoubtedly exist, but Options 1 through 8 are the major alternatives of interest. In the next section we will illustrate a comparison of these models using one commonly assumed scenario leading to binomial data. We will also consider model alternatives and their performance with two plausible scenarios giving rise to count data.

3. THREE EXAMPLES

In this section, three model-comparison scenarios are presented. The first illustration compares the models listed as Options 1 through 8 in the previous section with Binomial data. The second and third scenarios use count data arising in two different ways and compare models analogous to Options 1 through 8. All three scenarios assume a randomized complete block design with eight blocks and two treatments, referred to as treatment 0 (e.g. control or standard) and treatment 1 (experimental or test). Each scenario is divided into two parts: one with equal treatment means, the other with unequal treatment means. For each part, model comparisons are based on 2000 simulated experiments. Model-comparison criteria are

- rejection rate of $H_0 : \tau_0 = \tau_1$, the hypothesis of no treatment effect. For equal treatments, rejection rate implies type I error rate; for unequal treatments, rejection rate implies power.
- average estimate of the treatment mean on the data scale. For binomial data, this would be $\hat{\pi}_j$, $j = 0,1$, the estimate of the probability of a success for the i^{th} treatment – or the estimate of the marginal mean (which is *not* $\hat{\pi}_j$) for marginal models, as noted below.
- the average upper and lower 95% confidence bound for each treatment mean
- the observed percent coverage of the 95% confidence interval for each treatment mean

Finally, all ANOVAs, whether on untransformed or transformed data,were analyzed using standard REML-based LMM analysis assuming random block effect. The *conditional* GLMMs (i.e. those not defined on a quasi-likelihood) can be analyzed using either linearization or integral approximation. In this illustration,we compared linearization using SAS PROC GLIMMIX pseudo-likelihood (RSPL, the default, REML-like version) and integral approximation using adaptive quadrature. The Laplace integral approximation was also considered, but no results are shown because the difference between Laplace and adaptive quadrature was consistently negligible in the scenarios considered here (there are designs and distribution for which this would undoubtedly *not* be true). For *marginal* GLMMs (those defined on quasi-likelihoods) SAS GLIMMIX pseudo-likelihood was used. For GEE models (e.g. Option 7 shown above) true generalized estimating equation methods, e.g. SAS PROC GENMOD using a REPEATED statement or the GEE package in R, could have been used. In the interest of space, and because, for these scenarios, differences between GEE and pseudo-likelihood results are trivial, this was not done in these illustrations.

For R users, for the ANOVA models, REML results from the *lme4* package are identical to REML with SAS GLIMMIX. For conditional GLMMs, R's *lme4* and GLIMMIX adaptive quadrature yield essentially identical results. To the author's knowledge, R does not have a package equivalent to pseudo-likelihood; the closest is the GLMPQL package that implements penalized quasi-likelihood. However, GLMPQL imposes a scale parameter that cannot be turned off, meaning PQL cannot be used with a true conditional GLMM.

A. Binomial Scenario

This scenario used a beta-binomial template with Gaussian random block effects. There were four cases: all possible combinations of equal and unequal treatment probabilities with $N = 10$ and $N = 100$. Data generation was as follows.

- **Step 1.** The 8 block effects, $b_1, b_2, ..., b_8$ were generated from a $N\left(0, \sigma_B^2\right)$ distribution with $\sigma_B^2 = 0.5$
- **Step 2.**Block-perturbed, unit-level η_{ij} were computed for each experimental unit. In the equal treatment case, π_j was set to 0.9 for both treatments. Hence, $\eta_{ij} = \text{logit}\left(0.9\right) + b_i$ for $i = 1, 2, ..., 8;\ j = 0, 1$. In the unequal treatment case, when $N = 100$, $\pi_0 = 0.9$ and $\pi_1 = 0.8;$ when $N = 10$, $\pi_0 = 0.9$ and $\pi_1 = 0.7$. Hence, $\eta_{ij} = \text{logit}\left(\pi_j\right) + b_i$.
- **Step 3.** Block-perturbedprobabilities were computedas$1\Big/\left[1 + \exp\left(-\eta_{ij}\right)\right]$. Denote the block perturbed probabilities p_{ij}^{Blk}.
- **Step 4:** Probabilitieswere further perturbed at the unit-level according to a Beta distribution. Beta variates were generated using the fact that $Beta\left(c, d\right) =$ $\Gamma\left(c, 1\right)\Big/\left[\Gamma\left(c, 1\right) + \Gamma\left(d, 1\right)\right]$, where $\Gamma\left(c, 1\right)$ denotes a Gamma variate with shape parameter c and scale parameter 1, and expected value of the Beta variate is $c\big/\left(c + d\right)$. The Beta parameter d was set to 5 and c was obtained by solving $p_{ij}^{Blk} = c\big/\left(c + d\right) \Rightarrow c =$ $d\left[p_{ij}^{Blk}\Big/\left(1 - p_{ij}^{Blk}\right)\right]$. Thus, the unit level probabilities were generated as $\pi_{ij} \sim Beta\left(5\left[p_{ij}^{Blk}\Big/\left(1 - p_{ij}^{Blk}\right)\right], 5\right)$ variates.
- **Step 5.** The observed number of successes for the i^{th} block, j^{th} treatment was generated as a $y_{ij} \sim \text{Binomial}\left(N, \pi_{ij}\right)$ variate.

Notice that in this scenario, it is presumed that data arise by blocks and treatments affecting the likelihood of a success at the canonical parameter level, i.e. the level that corresponds to the model or link scale in a GLMM.

B. Count Scenarios

Count data have modeling options analogous to the binomial Options 1-8 described above. Denoting the observed count as y_{ij} the analogous options are as follows:

- **Option 1:** assume the counts are approximately normal. Fit the ANOVA model $y_{ij} = \mu + b_i + \tau_j + e_{ij}$, where $e_{ij} \sim NI\left(0, \sigma^2\right)$
- **Option 2:** use a variance stabilizing transformation. Three major types of transformations are commonly used for count data. Each has several variations. The following were used in this illustration
  - **Log:** $\log\left(y_{ij} + 1\right) = \mu + b_i + \tau_j + e_{ij}$ where $e_{ij} \sim NI\left(0, \sigma^2\right)$
  - **Square root:** $\sqrt{y_{ij} + 3/8} = \mu + b_i + \tau_j + e_{ij}$ where $e_{ij} \sim NI\left(0, \sigma^2\right)$
  - **Power:** $y_{ij}^{2/3} = \mu + b_i + \tau_j + e_{ij}$ where $e_{ij} \sim NI\left(0, \sigma^2\right)$

The log and square root transformations are forms suggested by Snedecor and Cochran (1989). The power transformation follows a suggestion from McCullach and Nelder (1989).

- **Option 3:** basic GLMM. $y_{ij} \mid b_i \sim \text{Poisson}\left(\lambda_{ij}\right)$; link and linear predictor $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + b_i + \tau_j$, $b_i \sim NI\left(0, \sigma_B^2\right)$

Like the binomial Option 3, the basic Poisson GLMM ignores unit-level perturbation of λ_{ij} and is, as a result, prone to overdispersion. Possibilities for accounting for unit-level perturbation include

- **Option 4:** add a block × treatment term to the linear predictor. The amended GLMM is $y_{ij} \mid b_i, (bt)_{ij} \sim \text{Poisson}\left(\lambda_{ij}\right)$; link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + b_i + \tau_j + (bt)_{ij}$, $b_i \sim NI\left(0, \sigma_B^2\right)$, $(bt)_{ij} \sim NI\left(0, \sigma_{BT}^2\right)$
- **Option 5:** reparameterize Option 4 as a compound symmetry model: $y_{ij} \mid (bt)_{ij} \sim \text{Poisson}\left(\lambda_{ij}\right)$; link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + \tau_j + (bt)_{ij}$, $\begin{pmatrix} (bt)_{i1} \\ (bt)_{i2} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_{CS}^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$. This is especially useful if Option 4 yields negative variance component estimates.
- **Option 6:** Add a unit-level scale parameter to the basic GLMM in Option 3. That is, $y_{ij} \mid b_i \sim quasi - \text{Poisson}\left(\lambda_{ij}\right)$; link $\eta_{ij} = \text{logit}\left(\pi_{ij}\right) = \eta + b_i + \tau_j$, $b_i \sim NI\left(0, \sigma_B^2\right)$, $quasi - \text{Var}\left(y_{ij} \mid b_i\right) \propto \phi\lambda_{ij}$
- **Option 7:**Replace the compound symmetry covariance structure in Option 5 with a working covariance. That is, $y_{ij} \sim quasi - \text{Poisson}\left(\lambda_j\right)$; link $\eta_j = \text{logit}\left(\pi_j\right) = \eta + \tau_j$, $quasi - \text{Var}\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} = \lambda_{ij}\phi \begin{bmatrix} 1 & \rho_w \\ \rho_w & 1 \end{bmatrix}$ where ρ_w denotes working correlation. Like the binomial Option 7, model targets the mean of the marginal distribution of y_{ij}, not λ_j, and is variously called a *marginal* GLMM or a GEE model.
- **Option 8:** Similar to Option 3 but assume $y_{ij} \sim \text{NegativeBinomial}\left(\lambda_{ij}, \varphi\right)$ rather than Poisson. This is actually Option 4 assuming thatthe rate parameter is perturbed by a Gamma-distributedunit-level random effect. Specifically, if $y_{ij} \mid b_i, u_{ij} \sim \text{Poisson}\left(\lambda_{ij} u_{ij}\right)$ and $u_{ij} \sim \Gamma\left(1/\varphi, \varphi\right)$ then $y_{ij} \mid b_i \sim \text{NegativeBinomial}\left(\lambda_{ij}, \varphi\right)$ One can think of the model as Poisson with link $\log\left(\lambda_{ij} u_{ij}\right)$ and linear predictor $\eta + b_i + \tau_j + \log\left(u_{ij}\right)$ where $u_{ij} \sim \Gamma\left(1/\varphi, \varphi\right)$, or as negative binomial with link $\log\left(\lambda_{ij}\right)$ and linear predictor $\eta_{ij} = \eta + b_i + \tau_j$ .In both cases, $b_i \sim NI\left(0, \sigma_B^2\right)$. The latter formulation meets strict GLMM requirements and can be fit using SAS PROC GLIMMIX. The former model can be fit, e.g. using SAS PROC MCMC. Both approaches yield essentially identical results.

Data Generation

Two scenarios were considered. The first simulates data arising fromblocks and treatments presumed to affect the expected counts at the canonical parameter level. The canonical parameter is $\log\left(\lambda_{ij}\right)$. The first scenario presumes that $\log\left(\lambda_{ij}\right)$ is a linear function of block and treatment effects. The second scenario presumes that the rate parameter itself is a direct linear function of block and treatment effects. These scenarios were compared to see if the data generation mechanism favors certain models. That is, GLMMs model block and treatment effects occurring on the link scale. If blocks and treatments affect the rate parameter as a linear function of $\log\left(\lambda_{ij}\right)$ this may give GLMMs an advantage relative to ANOVA models. On the other hand, ANOVA models presume that blocks and treatments affect the rate parameter as a direct linear function. Perhaps ANOVA models have an inherent advantage over GLMMs under scenario two.

In both scenarios, block effects were generated as Gamma variates so that the distribution of block effects would be right skewed. This allows investigating how the various models perform when the standard assumption of Gaussian random model effects is violated.

Scenario 1.

- **Step 1**. The 8 block effects, $b_1, b_2, ..., b_8$ were generated from a $\Gamma(5, 0.45)$ distribution.
- **Step 2**. Block perturbed canonical parameters, $\eta_{ij} = \eta_j + b_i$, were computed. For the equal treatments case, $\eta_j = \log(0.75)$. This implies $E\left(\eta_{ij}\right)$=1.96, which in turn implies that the expected, block perturbed rate parameter is $\exp\left[E\left(\eta_{ij}\right)\right] = 7$. Following the remaining steps yielded a marginal distribution of counts with similar median, mean and variance to a distribution resulting from Gaussian block effects and a rate parameter of 6.
- **Step 3**. Compute the block perturbed rate parameter, $\lambda_{ij} = \exp\left(\eta_{ij}\right)$
- **Step 4**. Generate the unit-level random effect, $u_{ij} \sim \Gamma\left(1/\varphi, \varphi\right)$. In this case, $\varphi = 0.5$.
- **Step 5**. Generate the observed counts as $\text{Poisson}\left(\lambda_{ij} u_{ij}\right)$ variates. Notice that this implies that, conditional on the block effects, the observed counts, denoted y_{ij} have a $\text{NB}\left(\lambda_{ij}, \varphi\right)$ distribution.

Scenario 1 results in count data with a right-skewed block effect distribution. Figure 1 shows the distribution of block effects generated in the equal treatment case. In the equal treatment case, the rate parameter – the expected value of $y_{ij} \mid b_i$ is the rate parameter $\lambda = 6$, and the marginal mean, $E\left(y_{ij}\right)$, is 14.5. For the unequal treatment case, in Step 2 $\eta_0 = \log(0.75)$ and $\eta_1 = \log(1.875)$ resulting in rate parameters of $\lambda_0 = 6$ and $\lambda_1 = 15$ and marginal means of 14.5 and 34.5 for treatments 0 and 1 respectively.

Scenario 2.

- **Step 1**. Generate 8 block-perturbed, pre-treatmentrate parameters, from a $\Gamma(6,1)$ distribution for the equal treatment case and from a $\Gamma(11,1)$ distribution for the unequal treatment case. Denote these values $\lambda + b_i$
- **Step 2**.For the equal treatment case, $\tau_0 = \tau_1 = 0$ and hence the result of Step 1 is the sum of the overall rate parameter plus the block effect. For unequal treatments, effects were set to $\tau_0 = -5$ and $\tau_1 = 5$. Block perturbed and treated rate parameters were thus computed as $\lambda_{ij} = \lambda + b_i + \tau_j$.
- **Step 3**. Generate the unit-level random effect, $u_{ij} \sim \Gamma(1/\varphi, \varphi)$. In this case, $\varphi = 0.5$.
- **Step 4**. Generate the observed counts as $\text{Poisson}(\lambda_{ij} u_{ij})$ variates. Notice that this implies that the observed counts, denoted y_{ij} have a $\text{NB}(\lambda_{ij}, \varphi)$ distribution.

Notice that in the equal treatment case, the expected rate parameter is 6. In the unequal treatments case, the expected rate parameters are 6 and 16 for treatment 0 and treatment 1 respectively. Also, in scenario 2 the rate parameter and marginal expectation of the observed counts are identical, whereas in scenario1, the marginal expected counts are considerably greater than the rate parameter.

4. SIMULATION RESULTS

Results for the binomial scenarios appear in Tables 1 through 4. Results for the count data scenarios appear in Tables 5 through 8. Before discussing these results, it is useful to review what we think we know about analyzing binomial and count data. What advice do students and practitioners get from current applied statistics literature? What is the reigning conventional wisdom among GEE and GLMM experts? Here is an admittedly non-exhaustive list of conventional wisdom articles of faith:

- **Convention Wisdom Item 1** (henceforth referred to as **CW1**): Standard ANOVA and GEE target the marginal mean. GLMMs target the expectation of $y_{ij} | b_i$, that is π for the binomial scenarios and λ for the count scenarios. Therefore, ANOVA and GEE should show more accurate confidence interval coverage for the marginal mean and GLMM should yield more accurate coverage of π and λ.
- **CW2**: As a consequence of the Central Limit Theorem, standard ANOVA should provide accurate type I error control and confidence interval coverage *in equal treatment case*. However, variance heterogeneity in the unequal treatment case conflicts with standard ANOVA's homoscedasticity assumption. If there are problems with standard ANOVA, they will show up in the unequal treatment case.
- **CW3**: Variance stabilizing transformations should address any problems with ANOVA in the unequal treatment case – assuming, that is, that transformations do not affect random model effects in ways that distort the resulting inference.

- **CW4:**GEEs are robust to model misspecification. If conventional wisdom is correct, GEEs should be less susceptible to violations of standard GLMM assumptions built into the scenarios considered in this paper, e.g. the beta-binomial process in Scenario 1 and the skewed block effect distributions in the count data scenarios.
- **CW5:**Sandwich estimators are widely regarded as preferable to model-based inferential statistics in conjunction with the GEE.
- **CW6:**Pseudo-likelihood has well-known accuracy issues with certain GLMMs. For example, Breslow and Clayton (1993), in their seminal GLMM publication, noted estimation bias issues with the logistic GLMM for binomial data with small cluster sizes. According to conventional wisdom, quadrature should address these issues.
- **CW7:**Following the last point, quadrature is generally portrayed as preferable to pseudo-likelihood. If this conventional wisdom turns out not to be true, the implications for R users are especially significant, since R's primary GLMM package, *lme4*, uses quadrature exclusively.

Each item on the above list follows from linear model theory, which is largely asymptoticfor GEEs and GLMMs. However, as experience with linear mixed models has shown, asymptotic theory does not always accurately predict small sample behavior. Surprises happen. As the sports pundits say (at these those not associated with college football's BCS), "The game is played on the field, not on paper." Simulation results below will be discussed in the light of how they support or, in several cases, contradict reigning GEE and GLMM conventional wisdom.

A. Binomial Data

Table 1 shows results for the equal treatment scenario with cluster size $N = 10$. Table 2 shows results for the equal treatment scenario with $N = 100$. Tables 3 and 4 show results for the unequal treatment scenarios with small($N = 10$) and large ($N = 100$) cluster sizes, respectively.In the tables, "analysis method" refers to the options listed in Section 2C.Each table shows the rejection rate (percent of simulated data sets for which $H_0 : \tau_0 = \tau_1$was rejected with $\alpha = 0.05$), the average estimate, average lower and upper 95% confidence bound, and the percent coverage for the 95% confidence interval. For standard ANOVA, the estimate is the mean of $p_{ij} = y_{ij} / N$ for each treatment. For all other methods, the estimate is the "Least Squares Mean" $\hat{\eta} + \hat{\tau}_j$expressed on the data scale: the back-transformation$\left[\sin\left(\hat{\eta} + \hat{\tau}_j\right)\right]^2$ for the arc sine square root transformation and the inverse logit link $1 \big/ \left[1 + \exp\left(\hat{\eta} + \hat{\tau}_j\right)\right]$ for the GEEs and GLMMs. Confidence bounds were obtained by first determining the 95% confidence interval for the least square means on the model scale then back-transforming or inverse-linking the lower and upper confidence bounds as appropriate. For the equal treatment case, estimates and confidence bounds were averaged over both treatments. For the unequal treatment case in Tables 3 and 4, estimates, confidence bounds and coverage are reported separately for each treatment. Also, in Tables 3 and 4, analysis methods that failed to control type I error, in other words, methods with rejection rates in excess of "nominal" as defined below, are shaded to indicate that these are methods one should avoid in practice.

In reporting the results, the word "nominal" will appear often. In the equal treatment case, the expected rate of type I errors is 0.05if $\alpha = 0.05$. Using standard margin-of-error methods for percent, with 2000 simulated experiments, “nominal” means a rejection rate between 0.035 and 0.065, with 0.03 or 0.07 characterized as “marginally nominal.” For confidence interval coverage,“nominal” means 0.935 to 0.0965, with 0.93 and 0.97 characterized as “marginally nominal.” For the unequal treatment case, the method that yielded maximum power among those methods whose type I error control was acceptable (i.e. nominal) was used as a reference for assessing the power of competing methods. The term “power loss” in the discussion below refers to how much lower the rejection rate of a given method was as a percentage of the reference method.

How did the methods perform with regard to conventional wisdom outlined above?

- **CW1**: ANOVA and GEE target the marginal mean; GLMMs target π.

ANOVA and GEE yielded average estimates equal to the marginal mean, 0.88 in the equal treatment case, 0.88 and 0.68 for unequal treatments with $N = 10$, 0.88 and 0.78 for $N = 100$. GLMMs all yielded average estimates within 0.01 of the target π for each scenario.

Confidence interval coverage was a different story. In the equal treatments case, ANOVA showed nominal confidence interval coverage of π for both cluster sizes (*N*=10 and *N*=100) but coverage for the marginal mean deteriorated as cluster size increased, yielding less than nominal coverage (0.92) when $N = 100$. For the unequal treatments case, ANOVA over-covered the larger mean (0.99) and under-covered the smaller mean (0.92) – as one would expect since ANOVA produces confidence intervals of equal width, whereas the variance of the binomial decreases as π approaches 1. With small cluster size, GEE provided nominal marginal mean coverage with the sandwich estimator but not with model-based statistics.The opposite occurred with larger cluster size: the sandwich estimate under-covered the mean (0.92) and yielded a greater than nominal type I error rate (0.086), whereas model-based statistics yielded acceptable results. SAS PROC GLIMMIX offers an optional bias correction for sandwich estimates due to Morel, et al. (2003). For the scenarios in this study, the Morel procedure was not needed with small cluster size and was excessively conservative with larger cluster size, yielding a power loss of 24%.

With small cluster size, the GLMM with unit-level random effectyielded higher than nominal confidence interval coverage (0.98). It did so in conjunction with extremely low rejection rate (type I error rate 0.008, power loss 12%) when implemented via pseudo-likelihood and unacceptably high rejection rate (type I error rate 0.154) when implemented via quadrature. Basic GLMM and GLMM with overdispersion scale parameter showed nominal coverage in the equal treatment case but greater than nominal coverage with unequal treatments. Only the GLMM assuming $y / N \sim Beta$ showed nominal coverage of π in both the equal and unequal treatment case.

With large cluster size, basic GLMM and the quadrature implementation of GLMM with unit-level random effect under-covered π and showed excessive type I error rate (0.215 and 0.113 respectively). The GLMM with overdispersion scale parameter, the pseudo-likelihood implemented GLMM with unit-level random effect and the GLMM assuming $y / N \sim Beta$ implemented either with quadrature or pseudo-likelihood showed nominal coverage of π and acceptable type I error control. The pseudo-likelihood implemented GLMM with unit-level random effect showed 5% power loss, but these four methods were otherwise quite similar.

- **CW2:** Standard ANOVA should perform nominally in equal treatment case but exhibit problems related to equal-variance assumption in unequal treatment case.

The only surprise here was that coverage of the marginal mean in the equal treatment case became worse as cluster size increased. Coverage was poor in the unequal treatment case regardless of cluster size.

- **CW3:** Transformation should address ANOVA issues in unequal treatment case.

Definitely not the case. The arc sine square root transformation is not needed when treatment means are equal and made things worse when treatment means were unequal. For small cluster sizes, confidence interval coverage was 0.89 and 0.93, both lower than nominal, for treatment 0 and 1 respectively. For large cluster size, coverage was 0.97 and 0.93, marginally higher than nominal for treatment 0 and marginally lower than nominal for treatment 1. These scenarios produced no evidence to support the use of the arc sine square root transformation in conjunction with mixed models and considerable reason to discourage their use.

- **CW4:** GEEs are robust to model misspecification.

No evidence from these scenarios. With small cluster size, the GEE with sandwich estimator yielded results equivalent to the GLMM with $y/N \sim Beta$ in terms of coverage, type I error control and power. Equivalent, but not better. With larger cluster size, the GEE with model-based statistics showed a 7% power loss. This is typical of GEEs with binomial data: because the marginal distribution is skewed toward 0.5, marginal means are closer together than corresponding π resulting in loss of power relative to the GLMM.

- **CW5:**Sandwich GEE estimators are preferable to model-based

No consistent evidence from these scenarios. Sandwich estimator yielded more accurate coverage and better type I error control for small cluster size. The opposite was true with large cluster size. In fact, it is well-known that sandwich estimators produce downward-biased standard errors and hence inflated type I error rates when the number of experimental units is relatively small, as they are in most agricultural research. The Morel correction was developed to address this bias, but it appears to be overkill for the scenarios investigated in this paper.

- **CW6:** Quadrature addressed well-known pseudo-likelihood issues in conjunction with small cluster size.

No evidence. For the GLMM with random unit-level effect, the GLMM one would expect to be most affected by cluster size, pseudo-likelihood did indeed perform poorly – confidence intervals coverage was well above nominal (0.98 and above) and the rejection rate was well below nominal (24-30% power loss). However, quadrature's performance was equally unacceptable, with type I error rate exceeding 0.15. Quadrature is a maximum likelihood procedure. It produces ML estimates of variance components and these have the same issues with GLMMs as their well-known problems with LMMs: downward-biased variance component estimates and hence inflated test statistics.

- **CW7:** Quadrature is preferable to pseudo-likelihood.

Not always. In these scenarios, not at all. See previous bullet point. This is of special relevance to R users. The primary GLMM package in R is *lme4*, which is a quadrature-only procedure. While SAS GLIMMIX was used for these simulations, for the GLMMs considered, *lme4* yields identical results. Pseudo-likelihood is superficially similar to penalized quasi-likelihood (PQL) , a procedure one can implement in R using the *glmpql* package. However, there is one crucial difference: PQL imposes a mandatory overdispersion scale parameter, whereas SAS GLIMMIX pseudo-likelihood does notinclude an overdispersion scale parameter on default –the user must specify it as an option. For the GLMM with unit-level random effect and the GLMM with $y/\ N \sim Beta$, adding an additional overdispersion scale parameter is nonsense. Until a true pseudo-likelihood package is developed in R, this is a major limitation on the literate use of GLMMs with R.

B. Count Data

Tables 5 and 6 show results for the equal and unequal treatments cases with negative binomial counts arising according to the link-scale mean process described in Section 3B Scenario 1. Tables 7 and 8 show results for the equal and unequal treatment cases with the additive mean model described in Section 3B, Scenario 2. Analysis methods correspond to the options described earlier in Section 3B. The rejection rate, average estimate, average lower and upper confidence bounds and confidence interval coverage for the rate parameter, λ and marginal mean are similar to criteria reported in the previous section for binomial data.

As with the binomial data, the estimates are of treatment sample means for standard ANOVA, back-transformed least square means for ANOVA with transformation, and inverse-linked least squares means for the GEEs and GLMMs. Three transformations, described under model Option 2 in Section 3C were considered. The inverse link for all GEEs and GLMMs is $\exp\left(\hat{\eta}+\hat{\tau}_j\right)$. One additional analysis method appears in Tables 5 and 7: the negative-binomial GLM – model Option 8 with fixed block effects. The inverse link for this model is $\exp\left(\hat{\eta}+\hat{\tau}_j+\left(1/8\right)\sum_i b_i\right)$. This was included to illustrate that the fixed-block/random-block decision, while inconsequential for complete block designs with Gaussian data, is *highly consequential* for GLMMs, even with complete block designs.

Link Mean Model Case

As with the binomial data, results are summarized with regard to ANOVA and GLMM conventional wisdom.

- **CW1:** ANOVA and GEE target the marginal mean; GLMMs target λ

Standard ANOVA and GEE models yielded average treatment mean estimates of 14.5, the marginal mean, for the equal treatment case, and 14.0-14.1 for treatment 0 and 34.6 for treatment 1, close to the marginal means of 14.5 and 34.5. While point estimates were accurate, confidence interval coverage was not, with coverage ranging between 0.69 and 0.84 compared to nominal 0.95.

The negative binomial GLMMs yielded accurate treatment mean estimates and nominal confidence interval coverage, as did the Poisson unit-level random effect GLMM provided it was implemented using pseudo-likelihood. When implemented with quadrature, the Poisson unit-level random effect GLMM showed inflated (0.125) type I error rate and less than nominal (0.85) coverage of λ. Basic GLMM yielded elevated type I error rate with scale parameter (0.195) and severely elevated type I error rate (0.459) with no overdispersion scale parameter.

Note that deciding whether to regard block effects as fixed or random matters. While the negative binomial GLMM with random block effects yields nominal results, the negative binomial GLM with fixed block effects yielded grossly inflated type I error (0.469) – similarto basic GLMM with no overdispersion parameter. In addition, confidence interval coverage (0.75) was substantially less than nominalalthough the point estimate itself was accurate (6.1 in the equal treatment case).

- **CW2:** Standard ANOVA should perform nominally in equal treatment case but exhibit problems related to equal-variance assumption in unequal treatment case.

In the equal treatment case, standard ANOVA yielded average treatment mean estimate of 14.5, equal to the marginal mean, but coverage was only 0.69. Interestingly, coverage was 0.95 for the rate parameter λ, not a result one would expect from standard ANOVA. However, the average confidence bounds were -5.1 and 34.1, the lower bound being especially unhelpful. What is the *actual* confidence associated with an interval that must be truncated at 0?In the unequal treatment case, ANOVA provided greater than nominal coverage of the treatment 0 mean (0.99) and lower than nominal (0.88) coverage of the treatment 1 mean. Also, for treatment 1, the average lower confidence bound was -18.7. Power loss was over 50%. Clearly, if this scenario accurately describes the process by which data arise, standard ANOVA will not do.

- **CW3:** Transformation should address ANOVA issues in unequal treatment case.

Mixed results.The log transformation yielded accurate estimates (6.2 in the equal treatment case, 6.1 and 15 in the unequal treatment case) and nominal confidence interval coverage. The square root and power transformation due to Nelder yield relatively inflated treatment mean estimates. Their confidence interval coverage was nominal in the equal treatment case but no better than standard ANOVA in the unequal treatment case. While the log transformation yielded accurate estimates and nominal type I error control, power loss was 16% relative to GLMMs that also provided nominal type I error control. Not evidence that supportsusing transformations.

- **CW4:** GEEs are robust to model misspecification.

No evidence from these scenarios. GEE with sandwich estimators performed poorly. The type I error rate was 0.191 with uncorrected sandwich estimators; power loss was over 50% with the Morel bias correction. With model-based statistics things were a little better but not much. Type I error rate was nominal, but power loss was over 30%. Confidence interval coverage was 0.78 for the marginal mean, 0.77 for λ.

- **CW5:**Sandwich GEE estimators are preferable to model-based.

No. See previous bullet.

- **CW6**: Quadrature is preferable to pseudo-likelihood.

Not always. The negative binomial GLMM yielded the most accurate analyses of any method considered: nominal type I error control, maximum power, nominal confidence interval coverage in both equal and unequal treatment cases. Pseudo-likelihood and quadrature yielded virtually identical results. On the other hand, the Poisson unit-level random effect GLMM yielded an analysis similar to the log transformation (accurate treatment mean estimates, good type I error control but underpowered) when implemented with pseudo-likelihood, but with quadrature, type I error rate was too high (0.125) and confidence interval coverage decreased to 0.85. This was among the approximately 80% of the data sets for which convergence could be obtained. The problem with quadrature is that it is uses numerical approximation to solve an exact likelihood. The Poisson GLMM assumes Gaussian block and unit-level effects. The data actually arose from skewed (Gamma) block effects and negative binomial unit-level data. When the discrepancy between model assumptions and data process is too great, quadrature struggles. Pseudo-likelihood seems to be more forgiving.

Additive Mean Model Case

Recall that the GLMM assumes a link-mean process and Gaussian block effects. In the link-mean scenario, only the block effect assumption was altered – their distribution was right-skewed Gamma rather than Gaussian. In the additive mean case both the mean model and block effect distribution violate GLMM assumptions. In theory, data arising from the additive mean model process should be closer to the standard ANOVA linear model, and ANOVA should have a competitive advantage relative to GLMM-based analysis. Also, if the GEE is truly more robust to model misspecification, this scenario should be revealing.

In fact, this scenario illustrates “the game is played on the field, not on paper” principle. While there are differences between the performance of the various analyses in the this scenario relative to the link-mean scenario, they are more differences in degree than in kind. What is truly striking is how little the big picture changes.

Standard ANOVA performs wellwhen treatment means are equal:nominal type I error rate, accurate confidence interval coverage. With unequal treatment means, however, ANOVA yielded poor confidence interval coverage (0.99 for treatment 0, 0.88 for treatment 1) and loss of power (20%). Transformations don’t help, but this time the Nelder power transformation is the least inaccurate and the log transformation is the most inaccurate.

GEE with sandwich estimates perform poorly, with or without Morel correction, for the same reasons they performed poorly in the link-mean case. As before, GEE with model-based statistics is a bit better, but in the unequal treatment case, confidence interval coverage is above nominal (0.97) for treatment 0, below nominal (0.92) for treatment 1, and power loss is just under 15%.

The basic GLMM still shows severely inflated type I error rate (0.26) with no overdispersion scale parameter. With scale parameter, the basic GLMM’s type I error rate is still inflated (0.082), but not quite as badly as in the link-mean scenario. The change of mean scenario does affect the negative binomial GLMM and the Poisson random-unit effect GLMM when implemented by pseudo-likelihood. For the negative binomial, type I error control and power are nominal but confidence interval coverage for treatment 0 is only 0.91; for the Poisson random unit-level GLMM power loss is similar to the GEE (just under 15%) and confidence interval coverage is below nominal (0.89 to 0.93). With quadrature, the model assumption failures for the Poisson unit-level random effect GLMM are so severe that convergence rate falls below 50% and theestimates one does obtain are nonsense. However the negative binomial GLMM implemented via quadrature yields the best performance of any analysis method – nominal type I error control, maximum power and generally nominal confidence interval coverage (marginally nominal, 0.93, for treatment 0 in the unequal treatment case. This is the only case in which quadrature did in fact outperform pseudo-likelihood.

As with the link-mean case, changing from random to fixed block effects dramatically alters results for the negative binomial GLMM. With fixed block effects, type I error rate was 0.209

and confidence interval coverage decreased to 0.78. Unlike ANOVA, with generalized linear models the fixed/random block effect decision is not one to make lightly.

5. CONCLUSIONS AND CAVEATS

First, the caveat. This paper should *not* be taken as a definitive, exhaustive comparison of analysis methods for binomial and count data. The results in the previous section apply to specific scenarios representing *examples* of how the probability processes giving rise to binomial and count data might be conceptualized and how the various alternatives for analysis perform under these scenarios. There are other plausible scenarios that, in the interest of time and space, were not considered in this paper.

The primary purpose of this paper is to illustrate how the ANOVA thought process can be adapted to construct plausible scenarios for how data arise consistent with the design structure, the primary response variable, and what is known – or what seems reasonable – aboutthe probability distribution of the design's major sources of variation. Students of statistical modeling and practitioners should find this useful. Students can use this approach to deepen their understanding of the theory and practice of statistical modeling. Practitioners can adapt the data generation and comparison of analysis alternatives demonstrated in this paper to conduct their own investigations to shed light on which method of analysis would be best suited to the needs to their particular problem.

While the results in Section 4 should not be considered exhaustive, they do provide food for thought both about how statistical scientists approach the analysis of non-normal data and what we teach beginning students and practitioners about analyzing non-normal data. The results here do call several items of common practice into question.

First, standard ANOVA's performance was *not*encouraging in any of the scenarios we investigated. As long as treatment means were *equal*, standard ANOVA typically provided nominal type I error control and confidence interval coverage. However, with unequal treatments ANOVA performed poorly. Power was reduced – in some cases drastically reduced –and confidence interval coverage of treatment means was inaccurate. Because the coverage issue stems from the mean-variance relationship in non-normal data, one would expect coverage inaccuracy to increase with increased difference among treatment means. Most agricultural experiments are conducted because researchers strongly suspect a treatment difference exists, so ANOVA's performance in the equal treatment case is something of a moot point. Often the question is not so much "is there a difference?" as it is "we know there is a difference – howbig is it?" The results here clearly cast doubt on whether ANOVA is up to the task of determining "how big?"

Second, transformations consistently were no help and often made matters worse. Almost everyone associated with agricultural research knows somebody whose opinion is, "why fool with that GLMM stuff? Just transform the data, compute ANOVA, and get on with your life." Introductory statistical methods typically "protect" students from generalized linear models by assuring them that transformation plus ANOVA is all they need. The results here should at least giveus pause. Researchers whose work depends on working with non-normal response variables are not well served by this kind of mentality.

Third, several items of GLMM conventional wisdom did not fare well. GEEs are reputed to be less susceptible to model misspecification. They showed no evidence of robustness in these scenarios. Integral approximation methods – quadrature and Laplace – are generally portrayed as preferable to pseudo-likelihood and penalized quasi-likelihood. These scenarios provided little evidence to support this. If anything, pseudo-likelihood appeared to be somewhat *more* robust than quadrature. Admittedly, these results are scenario dependent, but that is exactly the point. Sometimes quadrature is better, sometimes pseudo-likelihood is better, there is a place for both, and both options need to be understood and available. Finally, the models that did perform best in these scenarios are not the ones most likely to occur to users, nor are they necessarily alternatives that appear in current GLMM literature. For example, almost all GLMM literature focuses on the standard logit or probit binomial GLMM rather than the alternative with sample proportion distributed as beta.

The last two points have particular implications for two sets of GLMM users: the R community and Bayesians. R's only viable GLMM package, *lme4*, uses quadrature exclusively. This needs to change. In addition to *lme4*, R needs a flexible pseudo-likelihood package. The closest thing R currently has to a pseudo-likelihood package, *glmpql*, is antiquated, lacks needed features, and should not be considered a viable alternative. Bayesian methods and quadrature have in common the fact that they work from a well-defined and specific likelihood – a complex likelihood that can be evaluated only via numeric approximation or Monte Carlo methods or both, but a specific likelihood nonetheless. Several results in Section 4 illustrate how quadrature struggles and ceases to provide useable analysis when the model being fit differs substantially from the processes that gave rise to the data. Because Bayesian analysis is equally specific about the likelihood, it is similarly vulnerable. This is not so much a criticism of quadrature and Bayesian methods as it is a recognition that if these methods are to realize their potential, modelers must include an awareness of how data arise in their thought process and be disciplined about questioning whether the model they are fitting and the processes that gave rise to the data are a good match.

All this raises the final point. What do we tell students and practitioners? Here, we seem to be between a rock and a hard place. On one hand, standard ANOVA is clearly not a dependable tool for non-normal data. ANOVA with transformed data is even less suitable. ANOVA is especially problematic in the unequal treatment case, the case that, in practice, is usually the rule, not the exception. The power issue alone highlights the issue: few agricultural experiments are over-powered; inadequately replicated, under-powered experiments are far more common. University budgets being what they are – research budgets in general being what they are –we are not likely to see a profusion of lavishlyreplicated agricultural experiments in the foreseeable future. Given this reality, compounding the problem by using a method of analysis whose power is potentially less than 50% that of readily available, methodologically sound alternatives makes no sense.

On the other hand, the readily available, methodologically sound alternative, the GLMM, has a steep learning curve. The results in Section 4 suggest two things. On the good side, a well-chosen GLMM, properly implemented, provides nominal type I error control, accurate confidence interval coverage and maximum power. On the bad side, a poorly chosen GLMM, or a well-chosen but ineptly implemented GLMM, may yield results as bad as or even worse than standard ANOVA. How do we make the former – well-chosen, competently implemented GLMM-based analysis – available to agricultural researchers? What aspects of the GLMM are teachable – and *should* be taught – to consumers of statistical analysis?   What about graduate students and graduate curriculum in statistics programs? Currently, graduate programs are all over the map in terms of the amount of attention paid to the modeling-design-probability interface. Unfortunately, graduates from our programs with superficial, glib, a-little-knowledge-is-a-dangerous-thing semi-competence in this area are all too common. Part of the problem appears to be that the mind-set promoted by the traditional graduate-level modeling course, while well-suited to Gaussian data, does not help students to think more broadly. In fact, the traditional first course in modeling does much to misdirect and impede students' thinking about modeling non-normal data. This needs to change.

If a statistics graduate student or a linear models instructor asked, "What is the most important single thing in this paper?" I would tell them, without hesitation, focus on the WWFD process – learn to think clearly: how did the data arise?

6. REFERENCES

Bartlett, M.S. 1947. The use of transformations.*Biometrics***3**: 39-52.

Breslow, N.E. and D.G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**: 9-25.

Eisenhart, C. 1947. The assumptions underlying analysis of variance.*Biometrics***3**: 1-21.

Federer.W.T. 1955.*Experimental Design.* New York: MacMillan.

Ferrari, S.and F. Cribari-Neto. 2004. Beta regression for modeling rates and proportions. *J. Applied Statist.* **31**: 799-815.

Harville, D.A. 1976. Extensions of the Gauss-Markov theorem to include the estimation of random effects.*Annals of Statistics***4**: 384-395.

Harville, D.A. 1978. Maximum likelihood approaches to variance component estimation and to related problems. *J American Statist Assoc***72**: 320-338.

Henderson, C.R. 1953. Estimation of variance and covariance components.*Biometrics***9**: 226-252

Henderson, C.R. 1963. Selection index and expected genetic advance. in Hanson, W.D. and Robinson, H.F., *Statistical Genetics and Plant Breeding.* 141-163, *National Academy of Science - National Research Council Publication 982*

Laird, N.M. and J.H. Ware, 1982. Random-effects models for longitudinal data. *Biometrics***38**: 963-973.

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

McCullagh, P. and J.A Nelder.1989.*Generalized Linear Models, 2nded.* London: Chapman & Hall.

Miller.R.G. 1986.*Beyond ANOVA: Basics of Applied Statistics.* New York: Wiley.

Milliken, G.A. and D.E. Johnson. 2008. *Analysis of Messy Data, Vol. 1, 2nd Ed.* New York: Chapman and Hall.

Morel, J. G., M.C. Bokossa, and N.K.Neerchal. 2003. Small sample correction for the variance of GEE estimators.*Biometrical Journal*.**4**: 395–409.

Nelder, J.A. and R.W.M. Wedderburn. 1972. Generalized linear models. *J Royal Statist Soc A* **135**: 370-384.

S-189 Regional Project, Various Authors. 1989. *Applications of Mixed Models in Agriculture and Related Disciplines, SouthernCooperativeSeriesBulletinNo.343*. Baton Rouge, LA: LouisianaAgriculturalExperimentStation.

Snedecor, G. and W.G. Cochran. 1989. *Statistical Methods, 8th ed.* Ames, IA: Iowa State University Press.

Steel, R.D.G., J.H. Torrie and D.A. Dickey. 1980. *Principles and Procedures of Statistics: A Biometrical Approach 3rded.* New York: McGraw-Hill.

Stroup, W.W. 2013.*Generalized Linear Mixed Models.* Boca Raton, FL. CRC Press.

Yates, F. 1935. Complex Experiments. *J Royal Statistical Society, Supplement,* **2**: 181-223.

Yates. F. 1940. The recovery of inter-block information in balanced incomplete block designs. *Annals of Eugenics***10**: 317-325.

Figure 1. Distribution of Block Effects in Equal Treatment Count Data Scenario

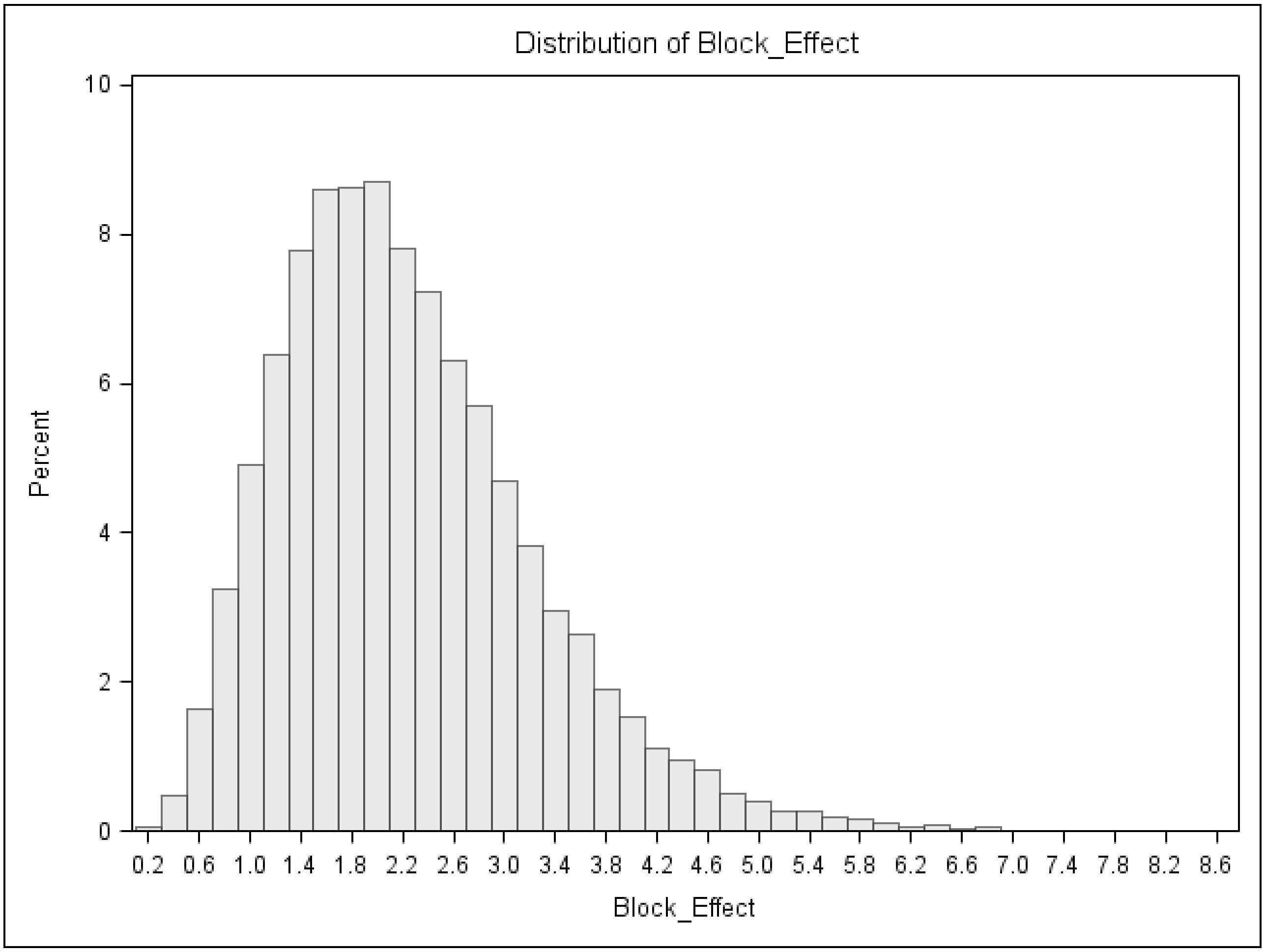

Table 1. Simulation Results, Binomial Response, Equal Treatment, N=10

Conditional $\pi_j = 0.9$; Marginal $\hat{\mu}_{P_j} = 0.88$; j=0,1

		average			coverage	
analysis method	rejection rate	estimate	Lower Confidence Bound	Upper Confidence Bound	conditional $\hat{\pi}_j$	marginal $\hat{\mu}_{P_j}$
standard ANOVA[1]	0.044	0.88	0.78	0.99	0.96	0.94
ANOVA with transformation[2]	0.051	0.92	0.80	0.99	0.90	0.83
GEE – model-based	0.027	0.88	0.73	0.95	0.96	0.98
GEE - sandwich	0.055	0.88	0.74	0.95	0.95	0.94
basic GLMM	0.021	0.90	0.75	0.96	0.96	>0.99
basic GLMM + scale parameter	0.052	0.90	0.75	0.96	0.96	0.97
GLMM + unit-level random effect – PL[3]	0.008	0.90	0.74	0.96	0.98	>0.99
GLMM + unit-level random effect – Q[4]	0.154	0.90	0.73	0.97	0.97	0.97
GLMM with unit ~ Beta – PL[3,5]	0.070	0.89	0.75	0.96	0.95	0.96
GLMM with unit ~ Beta – Q[4,5]	0.042	0.90	0.77	0.95	0.96	0.96

[1] response variable $p = y/n$

[2] response variable $\sin^{-1}\left(\sqrt{y/n}\right)$

[3] PL denotes “pseudo-likelihood” – SAS GLIMMIX Method=RSPL

[4] Q denotes adaptive quadrature – SAS GLIMMIX Method=QUAD

[5] response variable $p = y/n$, $p \mid b_i \sim Beta$

Table 2. Simulation Results, Binomial Response, Equal Treatment, N=100

Conditional $\pi_j = 0.9$; Marginal $\hat{\mu}_{P_j} = 0.88$; j=0,1

		average			coverage	
analysis method	rejection rate	estimate	Lower Confidence Bound	Upper Confidence Bound	conditional $\hat{\pi}_j$	marginal $\hat{\mu}_{P_j}$
standard ANOVA[1]	0.039	0.88	0.81	0.96	0.96	0.92
ANOVA with transformation[2]	0.048	0.90	0.82	0.95	0.94	0.88
GEE – model-based	0.051	0.88	0.78	0.94	0.94	0.95
GEE - sandwich	0.086	0.88	0.79	0.93	0.91	0.92
basic GLMM	0.215	0.90	0.82	0.95	0.93	0.86
basic GLMM + scale parameter	0.070	0.89	0.80	0.95	0.96	0.93
GLMM + unit-level random effect – PL[3]	0.048	0.90	0.82	0.95	0.95	0.89
GLMM + unit-level random effect – Q[4]	0.113	0.90	0.83	0.95	0.85	0.81
GLMM with unit ~ Beta – PL[3,5]	0.070	0.89	0.81	0.94	0.95	0.92
GLMM with unit ~ Beta – Q[4,5]	0.071	0.90	0.81	0.94	0.94	0.91

[1] response variable $p = y/n$

[2] response variable $\sin^{-1}\left(\sqrt{y/n}\right)$

[3] PL denotes “pseudo-likelihood” – SAS GLIMMIX Method=RSPL

[4] Q denotes adaptive quadrature – SAS GLIMMIX Method=QUAD

[5] response variable $p = y/n$, $p \mid b_i \sim Beta$

Table 3. Simulation Results, Binomial Response, Unequal Treatment, N=10

$\pi_0 \mid b_i = 0.9\,;\, \pi_1 \mid b_i = 0.7$

analysis method	reject rate	average for π_0				average for π_1			
		est	LCB	UCB	coverage	est	LCB	UCB	coverage
standard ANOVA[1]	0.617	0.88	0.73	1.03	0.99	0.68	0.53	0.84	0.92
ANOVA with transformation[2]	0.640	0.92	0.78	0.99	0.89	0.71	0.52	0.87	0.93
GEE – model-based	0.575	0.88	0.70	0.96	0.97	0.68	0.50	0.83	0.96
GEE - sandwich	0.661	0.88	0.74	0.95	0.95	0.68	0.50	0.83	0.94
GEE - swch+MBN[3]	0.464	0.88	0.68	0.96	0.98	0.68	0.45	0.85	0.99
basic GLMM	0.736	0.90	0.75	0.96	0.98	0.70	0.50	0.85	0.96
basic GLMM + scale parameter	0.647	0.89	0.74	0.96	0.97	0.70	0.49	0.85	0.97
GLMM + unit-level random effect – PL[4]	0.583	0.90	0.74	0.96	0.98	0.70	0.48	0.85	0.97
GLMM + unit-level random effect – Q[5]	0.670	0.90	0.73	0.96	0.95	0.71	0.49	0.85	0.94
GLMM with unit ~ Beta – PL[4,6]	0.661	0.89	0.74	0.96	0.96	0.70	0.49	0.84	0.94
GLMM with unit ~ Beta – Q[5,6]	0.582	0.88	0.73	0.95	0.96	0.70	0.52	0.84	0.94

[1] response variable $p = y/n$

[2] response variable $\sin^{-1}\left(\sqrt{y/n}\right)$

[3] sandwich estimate with small-sample bias correction using More1, et al. (2003) procedure

[4] PL denotes “pseudo-likelihood” – SAS GLIMMIX Method=RSPL

[5] Q denotes adaptive quadrature – SAS GLIMMIX Method=QUAD

[6] response variable $p = y/n$, $p \mid b_i \sim Beta$

Table 4. Simulation Results, Binomial Response, Unequal Treatment, N=100

$\pi_0 \mid b_i = 0.9\,;\, \pi_1 \mid b_i = 0.8$

	reject	average for π_0				average for π_1			
analysis method	rate	est	LCB	UCB	coverage	est	LCB	UCB	coverage
standard ANOVA[1]	0.585	0.88	0.78	0.98	0.99	0.78	0.68	0.88	0.92
ANOVA with transformation[2]	0.648	0.90	0.81	0.96	0.97	0.80	0.68	0.89	0.93
GEE – model-based	0.636	0.88	0.78	0.94	0.92	0.78	0.66	0.86	0.90
GEE - sandwich	0.712	0.88	0.79	0.94	0.89	0.78	0.68	0.87	0.92
GEE - swch+MBN[3]	0.519	0.88	0.76	0.95	0.97	0.78	0.61	0.89	0.97
basic GLMM	0.944	0.90	0.82	0.95	0.92	0.70	0.67	0.88	0.94
basic GLMM + scale parameter	0.707	0.90	0.81	0.95	0.94	0.79	0.66	0.86	0.94
GLMM + unit-level random effect – PL[3]	0.651	0.90	0.81	0.95	0.96	0.81	0.66	0.90	0.96
GLMM + unit-level random effect – Q[4]	0.738	0.91	0.82	0.95	0.86	0.81	0.68	0.89	0.85
GLMM with unit ~ Beta – PL[3,5]	0.685	0.89	0.80	0.95	0.95	0.70	0.66	0.88	0.95
GLMM with unit ~ Beta – Q[4,5]	0.678	0.89	0.80	0.95	0.93	0.70	0.67	0.88	0.94

[1] response variable $p = y/n$

[2] response variable $\sin^{-1}\left(\sqrt{y/n}\right)$

[3] sandwich estimate with small-sample bias correction using More1, et al. (2003) procedure

[4] PL denotes "pseudo-likelihood" – SAS GLIMMIX Method=RSPL

[5] Q denotes adaptive quadrature – SAS GLIMMIX Method=QUAD

[6] response variable $p = y/n$, $p \mid b_i \sim Beta$

Table 5. Simulation Results, Counts, Link-Scale Mean Model, Equal Treatment
Conditional $\lambda_j = 6$; Marginal $\hat{\mu}_{C_j} = 14.5$; j=0,1

			average			coverage	
analysis method		rejection rate	estimate	Lower Confidence Bound	Upper Confidence Bound	conditional $\hat{\lambda}_j$	marginal $\hat{\mu}_{C_j}$
standard ANOVA		0.028	14.5	-5.1	34.1	0.95	0.69
ANOVA on transformed response	$\log(c+1)$	0.053	6.2	2.0	17.3	0.94	0.49
	$\sqrt{c + 3/8}$	0.049	9.0	1.9	24.2	0.94	0.59
	$c^{2/3}$	0.043	10.5	1.6	27.4	0.95	0.63
GEE – model-based		0.058	14.5	4.6	62.0	0.78	0.77
GEE - sandwich		0.191	14.5	4.7	57.0	0.76	0.73
GEE –sandwich + MBN		0.075	14.5	3.5	85.4	0.89	0.84
basic GLMM		0.459	6.7	2.7	18.2	0.92	0.51
GLMM + OD scale		0.195	8.2	3.2	23.0	0.90	0.65
GLMM + unit-level random effect – PL		0.047	6.4	2.4	18.9	0.96	0.56
GLMM + unit-level random effect – Q		0.125	6.1	2.5	17.6	0.85	0.52
GLMM, $c \mid b_i \sim$ NB PL		0.057	7.2	2.7	20.9	0.94	0.63
GLMM, $c \mid b_i \sim$ NB Q		0.068	7.6	2.7	23.4	0.95	0.69
GLM, $c \sim \text{NB}$, fixed b_i		0.469	6.1	3.3	10.6	0.75	0.29

Table 6. Simulation Results, Counts, Link-Scale Mean Model, Unequal Treatment

Conditional: $\lambda_0 = 6\,;\lambda_1 = 15$ Marginal: $\mu_{C_0} = 14\,;\mu_{C_1} = 34.5$

		reject	average							
analysis method		rate	est	LCB	UCB	coverage	est	LCB	UCB	coverage
standard ANOVA		0.23	14.1	-18.7	46.8	0.99	34.6	1.9	67.4	0.88
ANOVA on transformed response	$\log(c+1)$	0.46	6.1	1.9	17.7	0.95	15.0	5.5	41.0	0.93
	$\sqrt{c+3/8}$	0.40	8.8	1.3	29.1	0.97	22.3	6.6	50.6	0.91
	$c^{2/3}$	0.35	10.2	0.6	34.6	0.98	25.8	6.7	56.0	0.90
GEE – model-based		0.37	14.0	3.4	94.9	0.87	34.6	13.4	105.8	0.70
GEE - sandwich		0.63	14.0	4.6	53.8	0.77	34.6	11.7	127.4	0.77
GEE –sandwich + MBN		0.40	14.0	3.1	92.3	0.92	34.6	9.3	175.7	0.87
basic GLMM		not shown – excessive type I error rate (see Table 5)								
GLMM + OD scale										
GLMM + unit-level random effect – PL		0.45	6.4	2.4	18.6	0.96	15.4	5.8	44.3	0.95
GLMM + unit-level random effect – Q		0.61	6.1	2.5	17.3	0.86	15.2	6.2	42.7	0.88
GLMM, $c\,|\,b_i$ ~ NB PL		0.55	7.2	2.7	20.5	0.94	17.6	6.8	49.7	0.95
GLMM, $c\,|\,b_i$ ~ NB Q		0.54	7.5	2.7	22.5	0.95	18.7	6.9	54.9	0.95

Table 7. Simulation Results, Counts, Additive Mean Model, Equal Treatment

$\lambda_j = 6$; j=0,1

			average			coverage
analysis method		rejection rate	estimate	Lower Confidence Bound	Upper Confidence Bound	conditional = marginal = $\hat{\lambda}_j$
standard ANOVA		0.038	6.0	1.5	10.5	0.95
ANOVA on transformed response	$\log(c+1)$	0.047	4.4	1.9	9.0	0.85
	$\sqrt{c + 3/8}$	0.047	5.1	2.0	9.4	0.89
	$c^{2/3}$	0.046	5.3	2.0	9.6	0.91
GEE – model-based		0.057	6.0	3.1	11.9	0.94
GEE - sandwich		0.087	6.0	3.1	12.0	0.93
GEE –sandwich + MBN		0.038	6.0	2.6	14.6	0.98
basic GLMM		0.260	5.1	2.8	9.4	0.88
GLMM + OD scale		0.082	5.6	2.8	11.6	0.95
GLMM + unit-level random effect – PL		0.049	4.9	2.5	10.0	0.92
GLMM + unit-level random effect – Q		0.283	4.6	3.1	8.1	0.58
GLMM, $c \mid b_i \sim$ NB PL		0.065	5.5	2.7	11.5	0.95
GLMM, $c \mid b_i \sim$ NB Q		0.061	5.6	2.7	11.8	0.94
GLM, $c \sim$ NB, fixed b_i		0.209	4.6	2.7	7.9	0.78

Table 8. Simulation Results, Counts, Additive Mean Model, Unequal Treatment

Conditional = Marginal: $\lambda_0 = 6\,;\, \lambda_1 = 16$

analysis method		reject rate	average							
			est	LCB	UCB	coverage	est	LCB	UCB	coverage
standard ANOVA		0.49	6.0	-1.2	13.2	0.99	16.1	8.8	23.3	0.86
ANOVA on transformed response	$\log(c+1)$	0.56	4.1	1.7	8.7	0.80	12.4	6.2	24.4	0.90
	$\sqrt{c+\frac{3}{8}}$	0.57	4.9	1.4	10.4	0.93	14.1	7.6	22.8	0.88
	$c^{2/3}$	0.55	5.1	1.1	11.2	0.96	14.7	8.0	22.8	0.87
GEE – model-based		0.53	6.0	2.7	14.1	0.97	16.1	9.7	27.0	0.92
GEE - sandwich		not shown – excessive type I error rate (see Table 7)								
GEE –sandwich + MBN		0.41	6.0	2.3	16.6	0.98	16.1	8.0	32.2	0.97
basic GLMM		not shown – excessive type I error rate (see Table 7)								
GLMM + OD scale		0.65	5.7	2.7	12.2	0.94	15.1	8.7	26.7	0.93
GLMM + unit-level random effect – PL		0.54	4.7	2.3	9.6	0.89	13.0	6.7	25.8	0.93
GLMM + unit-level random effect – Q		not shown: convergence failure rate > 50%								
GLMM, $c \mid b_i \sim$ NB PL		0.66	5.5	2.8	10.9	0.91	14.8	7.9	28.4	0.95
GLMM, $c \mid b_i \sim$ NB Q		0.62	5.6	2.7	12.0	0.93	15.2	7.8	30.2	0.96