Published February 19, 2014

Rethinking the Analysis of Non-Normal Data in Plant and Soil Science

Walter W. Stroup*

ABSTRACT

The introduction of high-quality, useable generalized linear mixed model (GLMM) software in the mid-2000s changed the conversation regarding the analysis of non-normal data from designed experiments. For well over half a century, the reigning paradigm called for using analysis of variance (ANOVA), either assuming approximate normality of the original data or applying a variance-stabilizing transformation. The appearance of GLMMs creates a dilemma. The ANOVA-based analyses and GLMM-based analyses often yield mutually contradictory results. What results should a researcher report, and how should the choice be justified? If GLMM-based analysis is preferred—and there is increasing evidence that this is the case—approaches to data analysis ingrained while learning ANOVA must be unlearned and relearned. The basic issues associated with the analysis of non-normal data are reviewed here, the thought processes required for GLMMs and how they differ from traditional ANOVA are introduced, and three examples are presented, giving an overview of GLMM-based analysis. The three examples include discussions of what is known to date about the relative merits of GLMM- and ANOVA-based analysis of non-normal data.

Through a manure trial on potato (*Solanum tuberosum* L.), Fisher (1923) introduced ANOVA. In the following years, ANOVA became institutionalized as the central feature of what is commonly accepted as standard statistical analysis for experimental research data. This understanding remains firmly in place today.

Analysis of variance rests on three assumptions: independent observations, normally distributed data, and homogeneous variance, the latter meaning that the variance among experimental units does not change with treatment. However, data with non-normal distributions are common in most areas of research. Examples include the percentage of seeds that germinate (binomial), weed count per plot (Poisson or negative binomial), time to flowering (exponential or gamma), disease rating category (multinomial), and proportion of leaf area affected (beta), to name a few. Data from emerging genomic and other "-omic" research often share characteristics with non-normal distributions. For all distributions except the normal, the variance depends on the mean. As a consequence, whenever the normality assumption is violated, the equal variance assumption must also be violated—at least, assuming that treatments affect the mean response. The question addressed here is: how should such data be analyzed?

Before 1990, this seemed to be a settled question. The Central Limit Theorem provided assurance that regardless of the distribution of the data, given a sufficient number of observations—read "properly designed experiment"—the sampling distribution of means could be assumed to be approximately normal. A considerable body of evidence for the robustness of ANOVA, summarized in an excellent overview by Miller (1997), accumulated during the 20th century. Standard variance-stabilizing transformations for common types of non-normal data were well known, included in statistical methods texts, and considered standard operating procedure in many agricultural disciplines.

Between the early 1990s and the late 2000s, advances in statistical theory and methodology that had been incubating for decades, enabled by rapid and sustained increases in computing capability, combined to dramatically change the conversation. The advance specifically relevant to this discussion is the GLMM. Generalized linear mixed models extend the linear model theory underpinning ANOVA to accommodate data that may be non-normal, may have heterogeneous variance, and, indeed, may be correlated. Viewed through the GLMM lens, the pre-1990s understanding of non-normal data—still pervasive in the agricultural research community—is antiquated at best, obsolete at worst.

Standard ANOVA on untransformed data, ANOVA with transformations, and GLMMs yield different, often contradictory and incompatible, analyses and conclusions, raising the question of what to report. Generalized linear mixed models require a change in mindset. Habits of mind acquired learning ANOVA apply essentially intact to transformed data but do not necessarily help, and often

Supplemental material available online. Department of Statistics, 340 Hardin Hall North, Univ. of Nebraska, Lincoln, NE 68583-0963. Received 16 July 2013. Accepted 20 Oct. 2013. *Corresponding author (wstroup1@unl.edu).

Published in Agron. J. 107:811–827 (2015)
doi:10.2134/agronj2013.0342

Available freely online through the author-supported open access option.
Copyright © 2015 by the American Society of Agronomy, 5585 Guilford Road, Madison, WI 53711. All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Abbreviations: AIC, Akaike information criterion; ANOVA, analysis of variance; GEE, generalized estimating equation; GLM, generalized linear model; GLMM, generalized linear mixed model; LM, linear model; LMM, linear mixed model; RCBD randomized complete block design; WWFD, What Would Fisher Do.

impede, working effectively with GLMMs. Those trained under the ANOVA paradigm typically find that GLMMs require considerable unlearning and relearning. Depending on the application, the GLMM learning curve can be steep. It is fair to ask the following: When will standard ANOVA give scientifically defensible, if less sophisticated, answers? Are transformations still relevant? When should the GLMM be considered *essential*? How does one know the difference? Answering these questions, as well as presenting an introduction to the thought processes and methodology of generalized linear models, is my primary objective here.

To help readers understand the paradigm shift occurring in applied statistics, I begin with a brief history. Fisher and Mackenzie (1923) published the first use of ANOVA for experimental data. Fisher (1925, 1935) established the template for statistics in agricultural research and rapidly became the template for experimental research in general. Yates (1940) introduced the recovery of interblock information—a crucial precursor to mixed model methodology—and was a leader in extending Fisher's work to complex experiments, notably split plots. Bartlett (1947) introduced transformations for non-normal data within the ANOVA framework. Eisenhart (1947), Henderson (1953, 1963), and Harville (1976, 1977) did seminal work essential to modern mixed models. Searle (1971) and Graybill (1976) integrated matrix algebra with linear model theory. This, along with the development of computers that could be programmed by written instructions, made modern statistical software (e.g., SAS) possible (SAS Institute, 2012). The SAS PROC GLM, a comprehensive linear model package, was introduced in 1976 and quickly became the standard ANOVA and regression package at North American agricultural research facilities. Despite its success, the limitations of PROC GLM and similar software were apparent from the outset. Specifically, the "general" linear model theory on which these packages were based were ill-suited to mixed-effects models—notably multilevel, split-plot type experiments—and non-normal data.

Nelder and Wedderburn (1972) introduced generalized linear models, a major departure in approaching non-normal data. Whereas transformations altered the data to meet ANOVA assumptions, Nelder and Wedderburn extended the linear model basis of ANOVA and regression to accommodate more plausible probability assumptions about the data. In 1982, the USDA-supported University Statisticians of Southern Experiment Stations, the group responsible for developing SAS, initiated a regional project to address PROC GLM's shortcomings. The project publication (University Statisticians of Southern Experiment Stations, 1989) along with Laird and Ware (1982) brought mixed model methods to the attention of larger research communities including agriculture. Before 1982, awareness of mixed models was confined to a few highly specialized applications. By the early 1990s, mixed model methods were mainstream. Liang and Zeger (1986) were similarly instrumental in expanding awareness and applicability of generalized linear models. In 1992, SAS introduced PROC MIXED, which implemented mixed model analysis for normally distributed data, and PROC GENMOD, which implemented fixed-effects-only generalized linear models for non-normal data. Breslow and Clayton (1993) and Wolfinger and O'Connell (1993) published seminal studies integrating mixed model and generalized linear model theory and methods. The next decade saw intense development of GLMM theory and methods. At the same time, computer technology was undergoing explosive development.

By the mid-2000s, practical GLMM software began to appear. The SAS PROC GLIMMIX was introduced in 2005. Several GLMM packages in R—GLMPQL, GEE, LME4, etc.—appeared as well. This was a watershed moment for statistical analysis. For the first time, useable software existed to implement the full range of statistical models explicitly intended to accommodate both complex experiments (primarily a mixed model issue) and non-normal data (primarily a generalized linear model issue). The extensive development of theory and methodology during the previous decades became available to researchers in accessible form. This explains why the entire question of how to do statistical analysis of non-normal data from experimental research is now being reassessed. Statistics, like all other disciplines, is dynamic; it is not a fixed set of unchanging rules passed down from Fisher and company.

I introduce the analysis of non-normal data using GLMMs, focusing on examples relevant to plant and soil science. I begin by presenting three motivating examples to illustrate the issues. I then provide an introduction to GLMM basics, especially the thought processes required to work effectively with GLMMs, followed by representative "how-to" examples. These examples are necessarily introductory in nature and are presented at survey- rather than textbook-level depth. A bottom-line summary, conclusions, and recommendations are provided. The examples, to the extent possible, focus on statistical issues and not software-specific programming details. Programs in R and in SAS with the data and basic statements for implementing the examples given here are available in the supplemental material.

MOTIVATING EXAMPLES

Why is ANOVA, with or without transformations, increasingly suspect as a tool for analyzing non-normal data, and why do GLMMs matter? The three examples in this section address these questions. Data from all three examples come from randomized complete block designs (RCBDs). Plant and soil scientists frequently use RCBDs, readers are familiar with them, and RCBDs illustrate many of the issues that arise in analyzing non-normal data. Each example has eight blocks, two treatments–generically called the "control" and the "test" treatment–and no missing data.

In the first example, the response variable is a count, e.g., the number of weeds in a plot. In the second example, the response variable is binomial—the number of observations with a characteristic of interest out of the total number of observations, e.g., the number of seeds that germinate out of 100 seeds per plot. The third example is a continuous proportion, e.g., the proportion of leaf area affected by a disease. Table 1 shows the data for these examples.

Count Example

Count data are discrete, non-negative, integer valued, and typically have right-skewed distributions. In classical probability theory, counts imply a Poisson distribution. As counts increase, the Poisson approximates the normal distribution—the larger the count, the better the approximation. Poisson variables have equal variance and mean. By definition, this means that Poisson counts violate the equal-variance assumption of ANOVA. For smaller counts, data analysts are often advised to use logarithmic

Table 1. Data for motivating examples.

	Count		Binomial†		Continuous proportion	
Block	Control	Test	Control	Test	Control	Test
1	1	36	98	94	0.573	0.925
2	5	109	95	36	0.044	0.835
3	21	30	93	85	0.888	0.949
4	7	48	94	88	0.008	0.941
5	2	0	99	91	0.990	0.994
6	6	2	61	82	0.409	0.958
7	0	5	84	43	0.117	0.520
8	19	26	92	71	0.926	0.975

† Number of events of interest per 100 observations.

or square root transformations. However, assuming Poisson-distributed counts is problematic for biological count data. Requiring the mean and variance to be equal is a rigid and usually unrealistic assumption. Based on biological theory and accumulated experience, the negative binomial provides a more realistic distribution. Negative binomial random variables have mean denoted $\lambda > 0$, and variance $\lambda + \phi\lambda^2$, where $\phi > 0$ is called the *scale parameter*. Notice that, unlike the normal, whose mean, denoted μ, and variance, denoted σ^2, are distinct entities, the negative binomial's variance depends on the mean and an additional scale term.

Table 2 shows the results of four analyses. The first is standard ANOVA with the count as the response variable. The second and third analyses use ANOVA on logarithmic and square root transformed counts, respectively. The logarithmic transformation uses the Snedecor and Cochran (1989) recommendation, log(count + 1). The square root transformation follows Schabenberger and Pierce (2002) and Kuehl (2000), who advise that $\sqrt{(\text{count} + 3/8)}$ is especially suitable for small counts. Although not shown here, one could also use $\text{count}^{2/3}$ as proposed by McCullagh and Nelder (1989). The fourth analysis is a GLMM that assumes that count has a negative binomial distribution.

The ANOVA on untransformed counts yields a p value of 0.0981 for the test of equal treatment means—marginal evidence of a treatment effect, but unconvincing in an "$\alpha = 0.05$" world. The treatment means are 7.6 and 32.0 for the control and test treatments, respectively. The standard error for both treatments is 9.1, which cannot possibly be true because the variance, and hence the standard errors, must be a function of the mean for count data, regardless of whether the distribution is Poisson, negative binomial, or any other plausible count distribution. The 95% confidence interval for the control treatment has, unhelpfully, a lower bound of –14. The confidence interval must therefore be truncated at

Table 2. Summary of analyses for motivating examples.

Treatment	Statistics	Standard ANOVA†	Transformation		Generalized linear mixed model
Counts (count ~ negative binomial)					
			log(count + 1)	$\sqrt{(\text{count} + 3/8)}$	
Control	mean	7.6	4.5	5.8	5.9
	SE(mean)	9.1	2.6	4.4	2.7
	95% confidence limits	–14, 22	0.8, 16.3	–0.2, 20.6	2.0, 17.5
Treated	mean	32.0	14.3	22.9	22.4
	SE(mean)	9.1	7.4	8.5	9.8
	95% confidence limits	10.4, 53.6	3.9, 46.7	7.1, 47.4	7.9, 63.3
Test of treatment mean difference	*F* value	3.64	3.33	4.32	7.17
	p value	0.0981	0.1107	0.0761	0.0316
Discrete proportion (successes ~ binomial)					
				$\sin^{-1}(\sqrt{\text{pct}})$	
Control	mean	0.90		0.92	0.93
	SE(mean)	0.062		0.040	0.030
	95% confidence limits	0.75, 1.04		0.81, 0.98	0.82, 0.97
Treated	mean	0.74		0.76	0.78
	SE(mean)	0.062		0.067	0.072
	95% confidence limits	0.59, 0.88		0.61, 0.88	0.57, 0.91
Test of treatment mean difference	*F* value	3.28		5.00	6.75
	p value	0.1132		0.0605	0.0355
Continuous proportion (proportion ~ beta)					
				$\sin^{-1}(\sqrt{\text{pct}})$	
Control	mean	0.49		0.49	0.49
	SE(mean)	0.11		0.20	0.11
	95% confidence limits	0.23, 0.75		0.18, 0.80	0.26, 0.72
Treated	mean	0.89		0.91	0.79
	SE(mean)	0.11		0.08	0.10
	95% confidence limits	0.63, 1.15		0.65, 1.00	0.48, 0.94
Test of treatment mean difference	*F* value	10.09		10.76	3.81
	p value	0.0156		0.0135	0.0919

† For counts: ANOVA directly on untransformed count data, assumes count ~ normal; for discrete proportion: ANOVA directly on pct = successes/100, assumes pct ~ normal; for continuous proportion: ANOVA directly on proportion, assumes proportion ~ normal.

0, raising the question, "Do we really have 95% confidence in a truncated interval"? These results strongly suggest the need for some alternative to the standard ANOVA.

With transformed counts, the test of equal treatment means yield p values of 0.1107 and 0.0761 for the logarithmic and square root, respectively. Stroup (2013a, 2013b) reported simulations that consistently showed the logarithmic transformation producing, on average, excessively conservative tests for count data—that is, less than the nominal α-level likelihood of rejecting the hypothesis of no treatment effect, and loss of power relative to statistically sound alternatives, notably the GLMM.

With the logarithmic transformation, the estimated counts are 4.5 ± 2.6 and 14.3 ± 7.4 for the control and test treatments, respectively. These are distinctly lower than the estimated counts from ANOVA on untransformed data. Again, this is typical of the logarithmic transformation. With the square root transformation, the estimates are 5.8 ± 4.4 and 22.9 ± 8.5. These are also lower than the untransformed ANOVA but not as low as the logarithmic transformation. Both transformations reflect increased variance with increasing count. As a consequence, for lower counts, confidence intervals for the mean are narrower than those produced by an untransformed ANOVA; for higher counts, the transformations result in wider confidence intervals. However, the logarithmic and square root transformations yield different confidence bounds. The logarithmic transformation precludes negative lower bounds, but the square root transformation does not.

The GLMM yields a p value of 0.0316, the only "significant" p value among the four analyses, assuming an "$\alpha = 0.05$" world. This obviously creates a dilemma. The estimated mean counts are 5.9 and 22.4, with standard errors of 2.7 and 9.8. The estimated counts are greater than those computed from the logarithmically transformed ANOVA, less than the standard ANOVA, and similar to the square root transformation. For lower counts, the confidence intervals are narrower than all three ANOVA-based methods. For larger counts, the GLMM-based confidence interval is shifted to the right but about the same width as its ANOVA-based analog.

These results are for a single data set, but they are not unique. They accurately represent the typical pattern of differences among these three approaches to analysis. Analysis of variance on untransformed data yields upwardly biased mean count estimates. The logarithmic transformation yields downwardly biased estimates. Square root transformations and GLMM analysis typically give unbiased mean count estimates, but the square root transformation does not necessarily yield sensible interval estimates. All four analyses produce similar p values when treatment means are roughly equal, but the GLMM yields lower p values than ANOVA, with or without transformation, when treatment differences exist. In other words, all four methods control Type I error adequately, but GLMMs have more power to detect treatment differences.

Discrete Proportions—Binomial Data

Discrete proportions arise from "yes–no" data—the plant is alive or dead, diseased or not, stalk lodging is present or it is not, the seed germinates or it does not. In each experimental unit, e.g., a plot, N observations are taken. For example, N plants are observed, and of these, Y show the response or characteristic of interest. Because N can vary among plots, most data analysts use the sample proportion, defined as pct = Y/N, as the response variable.

Formally, Y has a binomial distribution, written $Y \sim$ Binomial(N,p), where p denotes the probability that an observation drawn at random has the characteristic of interest. In multitreatment experiments, p_i denotes the probability for the ith treatment. Analysis focuses on estimating p_i for each treatment and testing the equality of p_i among treatments. The expected value of the binomial random variable is Np and its variance is $Np(1 - p)$. As with counts, the variance is a function of the mean; unlike either the normal distribution or the negative binomial distribution for counts, the binomial does not have a separate scale parameter. With the normal distribution, estimates of the mean and variance require distinct calculations; with the binomial, a single calculation, the estimate of p, determines both the mean and variance.

For large N, the normal distribution approximates the binomial; the approximation becomes more accurate as N increases. Textbooks give rules of thumb ranging from $Np \geq 5$ to $Np \geq 10$ if $P < 0.5$, or $N(1 - p) \geq 5$ to $N(1 - p) \geq 10$ if $p > 0.5$, although there is no universal agreement about what qualifies as "large." Even when the normal approximation is accurate, for multitreatment experiments, unequal p_i guarantees unequal variance, violating a key ANOVA assumption. The standard variance-stabilizing transformation for binomial data is the arc sine square root transformation, i.e., $\sin^{-1}(\sqrt{\text{pct}})$, also known as the *angular transformation*.

Now consider the three analyses. The p values are 0.1132, 0.0605, and 0.0355 for ANOVA on the untransformed pct, ANOVA on the transformed pct, and the GLMM, respectively. The estimated probabilities are 0.90, 0.92, and 0.93, respectively, for the control treatment and 0.74, 0.76, and 0.78 for the test treatment. The standard errors of the mean are equal for ANOVA on untransformed pct, which we know cannot be correct because the variance must change with changing p. Because test statistics depend on standard errors, this also invalidates the untransformed ANOVA p value. Both arc sine transformed ANOVA and the GLMM yield standard errors that reflect the mean–variance relationship.

As with the first example, these analyses produce incompatible results. For reasons explained below, there are two plausible understandings of p_i, called the *conditional* and the *marginal*. Each is appropriate for certain applications but not for others. Untransformed ANOVA yields estimates of the marginal p_i but not the correct standard errors; the GLMM yields estimates of the conditional p_i and correct standard errors. Arc sine transformed ANOVA does not provide estimates of either. Two issues for binomial data are (i) how does one decide which understanding of p_i applies in a given situation, and (ii) if the marginal p_i is appropriate, how does one obtain the correct estimate *and* appropriate standard errors and test statistics?

Continuous Proportions

Continuous proportions arise when a percentage is the response variable of interest but it does not arise from "Y out of N" binomial processes. Unlike normally distributed random variables, proportions are bounded from above, by 1, and below, by 0. When the mean proportion is close to 0 or 1, distributions tend to be skewed, whereas normality assumes a symmetric distribution. Processes giving rise to continuous proportions are best described by the beta distribution. The beta distribution has an expected value μ, where $0 < \mu < 1$ and variance $\mu(1 - \mu)/(1 + \varphi)$

Table 3. "What Would Fisher Do" ANOVAs for randomized complete block design.

Topographical		Treatment		Combined	
Source	df	Source	df	Source	df
Block	7			Block	7
		Treatment	1	Treatment	1
Unit(block)	8	"Parallels"	14	Unit(block)|treatment ("residual" or block × treatment)	8 – 1 = 7
Total	15	Total	15	Total	15

where $\varphi \geq 0$ and is referred to as the scale parameter. Unlike the scale parameter of the normal distribution, σ^2, the scale parameter of the beta distribution defines the variance only partially, not completely. Snedecor and Cochran (1989) stated that the arc sine square root or angular transformation can be used for continuous as well as binomial proportions.

For the three analyses, the *p* values are 0.0156, 0.0135, and 0.0919 for ANOVA on the untransformed pct, ANOVA on the angular transformed pct, and the GLMM assuming a beta distribution, respectively. The estimated proportions for the control treatment are 0.49 ± 0.11, 0.49 ± 0.20, and 0.49 ± 0.11, respectively, with confidence intervals (0.23,0.75), (0.18,0.80), and (0.26,0.72). For the test treatment, the estimates are 0.89 ± 0.11, 0.91 ± 0.08, and 0.79 ± 0.10. The confidence interval for the test treatment mean obtained from untransformed ANOVA extends from 0.63 to 1.15—effectively this means (0.63,1) because proportions cannot exceed 1. As with the analyses above, this raises a question about the true confidence level that can be legitimately attached to truncated intervals. For proportions close to 0.5, all three methods yield similar estimates of the mean, but the angular transformation's confidence interval is much wider than that of the untransformed ANOVA and the GLMM. For proportions close to 0 or 1, the angular transformation shows a greater impact of changes in proportion on standard errors, whereas the GLMM yields estimated proportions that appear to be attenuated toward 0.5.

Unlike the count and binomial examples, the *p* values from ANOVA are lower—"more significant"—than the GLMM. Once again, however, the results are contradictory. Which analysis should we report?

CONTEMPORARY METHODS FOR NON-NORMAL DATA—GENERALIZED LINEAR MIXED MODELS

Analysis begins with a statistical model, a description of the impact of experimental factors and random variation on the observed response. For designed experiments, the statistical model and the ANOVA table are intimately linked. The ANOVA table is a good place to start to understand GLMMs, how they differ from traditional ANOVA models, and how to set up and work with GLMMs.

Stroup (2013a) introduced an exercise called "What Would Fisher Do?" (WWFD) to help students and data analysts work through the steps leading from a description of the design to the ANOVA table to the model. The procedure was based on Fisher's comments following Yates (1935). Fisher said that any experiment could be described in terms of its "topographical" and "treatment" components, the former being the physical elements such as blocks or experimental units—what Federer (1955) and later Milliken and Johnson (2009) would call the "experiment design." By writing an ANOVA for the topographical design, another for the treatment design, then integrating them, Fisher suggested that the appropriate analysis would then be apparent. Stroup's WWFD adapted Fisher's thought process by showing how the integrated ANOVA is translated into a GLMM.

To illustrate, consider the RCBD from the examples above. The design had eight blocks, with two experimental units per block. The two treatments were randomly assigned to experimental units, one unit per block per treatment. Thus the sources of variation for the topographical ANOVA are blocks and units within blocks; the sources of variation for the treatment ANOVA are treatment and whatever is left, a term Fisher called *parallels*. Table 3 shows the topographical, treatment, and combined ANOVA tables.

The placement of the rows in the topographic and treatment ANOVAs is important. Sources of variation in the treatment ANOVA are always placed in the line immediately above the line in the topographical ANOVA corresponding to the unit to which they were applied. Here, "treatment" was randomly assigned to "unit(block)." Then one "slides" the sources of variation to the right to obtain the combined ANOVA. Notice the name of the source of variation in the last line of the combined ANOVA: unit(block)|treatment. Read this as "unit within block after accounting for treatment." Also notice that its seven df result from the original df for experimental units in the topographical ANOVA minus the df of the treatments applied to those units. While these line placement and df protocols seem obvious with a simple, single-factor design such as the RCBD, disciplined application of WWFD rules greatly facilitates defining sensible models for arbitrarily complex designs.

Traditionally, this leads to an equation read as "observation = overall mean + treatment + block + error" and written in statistical notation as $y_{ij} = \mu + \tau_i + b_j + e_{ij}$, where y_{ij} denotes the observation on the *i*th treatment and *j*th block, and each line in the combined ANOVA implies a corresponding term on the right-hand side: "block" implies b_j; "treatment" implies τ_i; and "unit(block)|treatment" implies e_{ij}. Statistical texts commonly refer to e_{ij} as the *residual* or *random error* term, assumed to be independent and normally distributed, with mean 0 and variance σ^2. This equation provides the standard justification for using the tools of ANOVA—sums of squares, mean squares, *F* tests, etc.—provided, that is, that the normality assumption is plausible. Without normality, however, as we saw above, we have a problem.

The GLMM takes a different approach. Rather than a single model equation, the GLMM defines two terms: the *linear predictor* and the *distribution* of the observations at the unit level, i.e., at the level where measurements are taken. For example, for normally distributed data, write the distribution of the observations as $y_{ij} \sim N(\mu_{ij}, \sigma^2)$, read "the observations (y_{ij}) have a normal distribution with mean μ_{ij} and variance σ^2." Write the linear predictor as $\mu_{ij} = \mu + \tau_i + b_j$. Notice that the linear predictor describes how the observation's mean is affected by the sources of variation, block,

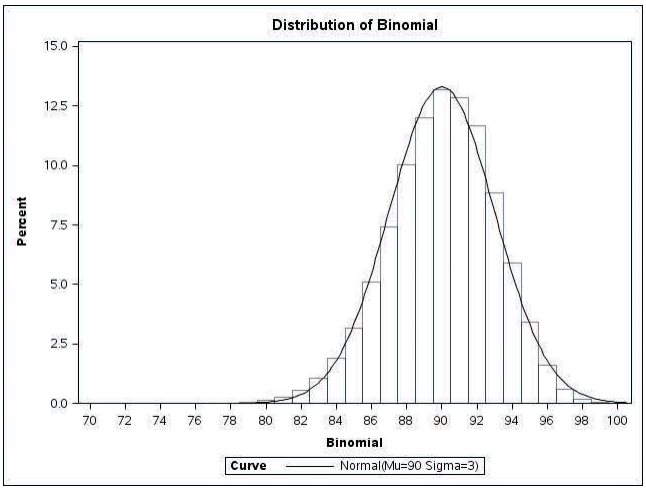

Fig. 1. Binomial probability distribution, N = 100, p = 0.9, with normal distribution superimposed.

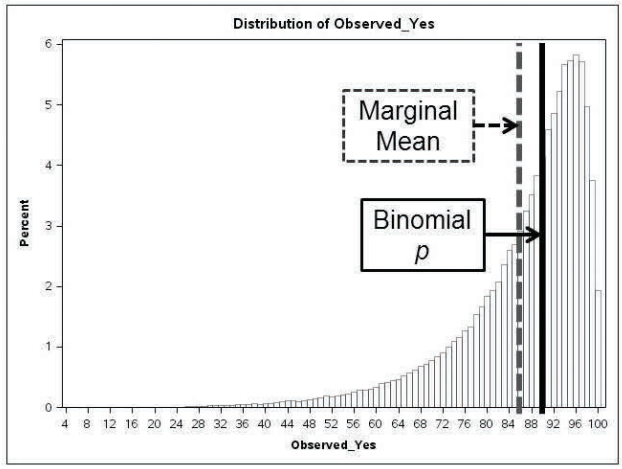

Fig. 2. Marginal distribution = distribution of observed number of successes from randomized complete block designs with binomial data.

and treatment, listed in the ANOVA. This is called the *probability distribution* formulation of the model; the traditional approach in the previous paragraph is called the *model equation* formulation.

For normally distributed data, these two formulations are simply different ways of expressing the same thing. Readers may ask, "If it doesn't matter, why add a new complication?" The answer: because the probability distribution approach accommodates non-normal data *competently*, whereas the model equation does not.

To see this, we need a brief excursion into probability and estimation theory. The following is intended as a "what consumers need to know" guide. For the rigorous, "gloves off" presentation of GLMM theory, see Stroup (2013a). For most agronomists, full immersion in GLMM theory is unrealistic and unnecessary. What they do need is a general understanding of the main issues and how contemporary statistical science approaches them. Some GLMM intuition is essential to appropriately analyze non-normal data and interpret the results.

To illustrate, consider the binomial example from above. The observations made on each experimental unit—hereafter referred to as the *ij*th unit: the *i*th treatment in the *j*th block—are assumed to have a binomial distribution with N_{ij} "yes–no" observations (e.g., seeds that either do or do not germinate) and probability p_{ij} of a "yes" response (e.g., the seed germinates) on any given observation. In the binomial example above, $N_{ij} = 100$ for all experimental units, allowing us to replace N_{ij} with N. Thus the mean for the *ij*th unit is Np_{ij} and the variance is $Np_{ij}(1 - p_{ij})$. Formally, the distribution is denoted Binomial(N,p_{ij}). Figure 1 shows an example binomial distribution with $p_{ij} = 0.9$. Figure 1 also shows the normal distribution with mean $Np_{ij} = 100 \times 0.9 = 90$ and standard deviation $[100 \times 0.9 \times (1 - 0.9)]^{1/2} = 3$ superimposed.

Inspection of Fig. 1 reveals why, with large N, the normal approximation to the binomial reassures and tempts data analysts. However, the reassurance is only apparent. Life is not so simple. The WWFD ANOVA tells us that we have two sources of variation, block and treatment, that affect p_{ij}. First consider the block. Blocking is a design strategy to ensure that units within blocks are as similar as possible. Variability *among* blocks is expected; variability *within* blocks is minimized to the extent possible. In addition, there is nothing special about the blocks actually used in an experiment. Any set of blocks with similar characteristics will do. In other words, we assume that the blocks are representative of blocks we could have used. Taken together, *representative* and *variability* strongly suggest *sample* and *probability distribution*. Typically, variation among blocks is assumed to follow a normal distribution.

What happens when p_{ij} is affected by a block effect that has a normal distribution? Figure 2 shows the distribution of the resulting observations. This is the distribution of the data that we actually observe. It has two features that are crucial to understanding the logic of analyzing non-normal data.

First, even though the normal approximation is convincing for the binomial distribution at the *ij*th unit level, and the distribution of the block effects is normal by assumption, the resulting distribution of the observations is not even remotely normal. In this case it is strongly left-skewed. In general, for $p_{ij} > 0.5$, the distribution of the observations will be left-skewed; for $p_{ij} < 0.5$, the distribution will be right-skewed. The skewness increases as the probability approaches 0 or 1. The observations are symmetrically distributed *only* when $p_{ij} = 0.5$ for all treatments—not probable in practice.

The second important feature of the distribution shown in Fig. 2 is that its mean is 86.6. Consequently, the mean sample proportion is 0.866. This is important because standard ANOVA—the model equation approach defined above—yields an unbiased estimate of the mean of the distribution of the observations. If your goal is to estimate the binomial probability, $p_{ij} = 0.9$, you will not get it using ANOVA. Analysis of variance will give you an unbiased estimate of 0.866. A similar set of illustrations could be developed for the count data above or the continuous proportion data.

Formally, there are three distributions relevant to the analysis of experimental data. Two of them follow from sources of variation identified by the WWFD ANOVA exercise: the distribution at the unit level (the binomial distribution in our example) and the distribution of effects that are considered "representative" and having "variation" (blocks in our example). The latter are known as *random effects* in statistical modeling. We refer to the latter as the distribution of the random effects and the former as the distribution of the observations conditional on the random effects. In formal statistical notation, write the block distribution as $b_j \sim NI(0,\sigma_B^2)$, read "block effects are normally and independently distributed with mean 0 and variance σ_B^2." Write the conditional distribution of unit-level observations as $y_{ij}|b_j \sim$ Binomial(N,p_{ij}),

read "the distribution of the observations, conditional on the observation being in the *j*th block, is Binomial with N yes–no observations per experimental unit and probability p_{ij} of a 'yes' response for any given observation."

The block and unit-given-block distributions follow from the sources of variation listed in the ANOVA; however, neither can be observed directly. The only distribution we can actually observe is called the *marginal distribution* of the observations. The important thing for users to remember is that when we say we have "binomial" data, we are referring to the distribution of the observations conditional on the *ij*th unit. The distribution of the observed data, however—the marginal distribution—is emphatically not binomial.

The reason this is a non-issue with normally distributed data is that if the random-effects distribution and the unit-given-random-effects distribution are normal, the resulting marginal distribution is normal as well. This only happens, however, when the unit-of-observation level distribution, that is, the distribution of $y_{ij}|b_j$, is normal. For all other data—binomial, counts, continuous proportions, time-to-event, etc.—the marginal distribution of the observed data is quite different. Our usual intuitions can betray and mislead.

The fundamental problem of analyzing non-normal data, especially with the designs most commonly used in agronomic research, is that what we want to estimate or test—e.g., treatment effects on p_{ij} for binomial data—involves parameters of distributions we cannot directly observe. What GLMM analysis does that standard ANOVA and regression cannot do is provide a way to extract the information we want—about the effects listed from the WWFD ANOVA exercise— from the observations we have, where these effects are camouflaged in a complex marginal distribution.

This ends the first excursion into probability. Returning to the task of writing a GLMM for binomial data from a randomized block design, we need to specify linear predictor and probability distributions. The distributions we have already specified: $b_j \sim NI(0,\sigma_B^2)$ for blocks, $y_{ij}|b_j \sim$ Binomial(N,p_{ij}) for the observations conditional on the blocks. Just as for normally distributed data, the linear predictor must account for treatment and block. A logical candidate is $\eta_{ij} = \eta + \tau_i + b_j$. The form is similar to its counterpart for normal data, but there are small but important differences. The left-hand side of the equation is not μ_{ij}, and because η_{ij} is not the mean, we also replace μ (the overall mean) by η (the intercept). We use η_{ij} and not the mean because, with non-normal data, we get better accuracy from analyses in which the treatment and block affect the mean indirectly, not as a direct additive equation.

To see why, we need to make another brief detour into probability. The probability of getting y_{ij} "yes" out of N observations is given by the formula

$$\binom{N}{y_{ij}} p_{ij}^{\,y_{ij}} \left(1-p_{ij}\right)^{N-y_{ij}}$$

This is called the binomial *distribution function*. Estimation theory uses the natural logarithm of the distribution, called the *log likelihood*. The binomial log likelihood is

$$\log\binom{N}{y_{ij}} + y_{ij} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) + N \log\left(1-p_{ij}\right)$$

The term $y_{ij}\log[p_{ij}/(1-p_{ij})]$ is particularly important: in the log likelihood, whatever is multiplied by y_{ij} is called the *natural* (or *canonical*) *parameter*. The natural parameter is always a function of the mean. For the binomial, the mean is Np_{ij} and the natural parameter is $\log[p_{ij}/(1-p_{ij})]$. In categorical data, $\log[p_{ij}/(1-p_{ij})]$ is also called the *log odds*. In modeling theory, it is called the *logit*.

Why is this important? Without belaboring the underlying theory, the natural parameter is a better candidate for regression and ANOVA-like models that the mean itself. Following this idea, we write the linear predictor as $\eta_{ij} = \log[p_{ij}/(1-p_{ij})] = \eta + \tau_i + b_j$. In generalized linear model terminology, η_{ij} is called the *link function*. It is the function that "links" the mean to the linear predictor. The *inverse link* expresses the mean in terms of the linear predictor. For the logit, the inverse link is $p_{ij} = 1/[1 + \exp(-\eta_{ij})]$.

Note that the link function does not have to be the natural parameter. In some cases, other functions of the mean make more sense. Nonetheless, the natural parameter is a good way to think about the logic of the link function and the linear predictor. Also, more often than not, the natural parameter does serve as the link function. See Table 4 for a list of commonly used link functions for distributions of interest in the plant and related sciences.

To summarize, full specification of the linear model requires, at a minimum, three elements:

- the unit-level distribution
- the linear predictor
- the link function

If the linear predictor contains random effects, then a fourth element, the distribution of the random model effects—or distributions if there is more than one random effect—must be specified. In this case, the unit-level distribution is understood as the conditional distribution of the observations given the random effects.

Table 4. Mean, variance, and usual link function for common distributions.

Distribution	Mean	Variance	Link
Normal	μ	σ^2	μ (identity)
Negative binomial	λ	$\lambda + \phi\lambda^2$	$\log(\lambda)$
Poisson	λ	λ	$\log(\lambda)$
Binomial proportion	p	$p(1-p)/N$	logit = $\log[p/(1-p)]$† probit = $\Phi^{-1}(p)$
Beta	μ	$\mu(1-\mu)/(1+\varphi)$	$\log[\mu/(1-\mu)]$
Exponential	μ	μ^2	$\log(\mu)$‡
Gamma	μ	μ/φ	$\log(\mu)$‡

† Natural parameter is logit.

‡ Natural parameter is $1/\mu$ but rarely used as link function; log is a better choice.

Table 5. Linear model (LM), linear mixed model (LMM), generalized linear model (GLM), and generalized linear mixed model (GLLM) classified by common response variable types in conjunction with effects needed in the model.

Response type	Example response variables	Distribution	Fixed effects		Mixed effects	
			Categorical	Continuous	Repeated measures in time or space	Random block, split plot, etc.
Continuous symmetric	height, weight, yield	normal	LM		LMM	
Count	number of weeds, insects	negative binomial, Poisson	GLM		GLMM	
Discrete proportion	*Y* "yes" out of *N* observations	binomial				
Continuous proportion	leaf area percentage	beta				
Continuous, nonzero, skewed	time to event	exponential, gamma				

When the unit-level distribution is normal, the link function is often not mentioned because it is equal to the unit-level mean. This matters to agronomic researchers mainly when describing the statistical analysis used, for example in the Materials and Methods section of a thesis or journal article. If the data are normal, the identity link is understood; if the data are non-normal, the link should be identified.

In contemporary statistical terminology, any model so specified is technically a GLMM; however, linear models are usually referred to by the following acronyms:

- LM: linear model—unit-level distribution is normal; fixed-effect-only model, that is, no random model effects.
- LMM: linear mixed model—unit-level-given-random-effects distribution is normal; linear predictor contains random model effects. Blocked designs where block effects are considered random (see discussion above) and all designs with split-plot features require random model effects; all repeated measures in time or space require mixed model methodology.
- GLM: generalized linear model—unit-level distribution is non-normal; fixed-effects-only linear predictor.
- GLMM: generalized linear mixed model—unit-level-given-random-effects distribution is non-normal; linear predictor contains random effects.

Table 5 summarizes response variables by model type combinations common in plant and soil science research. This table can be used to help organize thinking about aligning design and data to an appropriate model.

Before moving on, a comment about the acronym GLM is in order, because it is guaranteed to confuse researchers raised on ANOVA via SAS. From its introduction in 1976 through the mid-1990s, PROC GLM was the preeminent SAS procedure for analyzing experimental data. At the time it was introduced, GLM meant "general linear model." Don't confuse *general* and *generalized*—they have very different meanings in statistical science. The general linear model is "general" in the sense that it can accommodate any linear predictor, provided the data are normally distributed and all model effects are fixed. By 1976 standards, that was "general." By 2014 standards it is not. The general linear model cannot fully accommodate random model effects nor can it accommodate non-normal data (except via transformations, which do not *accommodate* non-normal data so much as they attempt to *force* data to *act normally* instead). The contemporary acronym for the general linear model is LM. The SAS PROC GLM can compute the analysis of LMs but not GLMs.

In SAS, PROC MIXED is specifically intended for LMMs. Many researchers still use PROC GLM for certain mixed models but this is emphatically discouraged—see Littell et al. (2006) for a full treatment of this subject. The GLMM integrates mixed and generalized linear models. The PROC GLIMMIX procedure was developed for GLMMs. Notice that all linear models are special cases of the GLMM: the GLM is a GLMM with no random model effects; the LMM is a GLMM with normally distributed data; the LM is a GLMM with no random effects and normally distributed data. For this reason, software intended to implement GLMMs can implement any linear model. The PROC GLIMMIX can compute simple ANOVA—using essentially the same syntax and yielding the same statistics relevant to statistical inference. For most data analyses, PROC GLIMMIX has effectively replaced PROCs GLM, MIXED, and GENMOD. On the other hand, PROC GLM cannot implement LMMs, GLMs, or GLMMs.

Generalized linear models and GLMMs raise three crucial issues that are likely to be unfamiliar to those new to generalized models. These are

- the model scale vs. data scale
- what terms to include in the linear predictor and what terms must not be included
- conditional vs. marginal inference and, as a consequence, conditional vs. marginal models

These are discussed briefly below. They are developed in greater detail and with more relevant context in the examples below. Interested readers are referred to Gbur et al. (2012) and Stroup (2013a) for more in-depth discussions of these issues.

Model vs. Data Scale

For GLMs and GLMMs, all statistical analysis occurs in terms of the link function. For example, with the binomial example above, analysis is in terms of the logit. The test of equal treatments implies testing for equal τ_i and hence equal logits. Ask for the "mean" and you get the mean logit. This is fine for testing—tests using GLMMs are demonstrably more accurate than standard ANOVA, with or without transformation—but the mean logit or the difference between two mean logits is not the stuff of understandable reports. Instead, an understandable report of the results ought to include estimates of the probabilities, p_i, for each treatment, obtained using the inverse link. Estimates of the mean logit are examples of the *link* or *model* scale. The estimate converted to a probability is an example of the *data* scale.

Notice that with normally distributed data, there is no data scale–model scale distinction. The treatment mean estimate

computed in terms of the model is in fact an estimate of the ith treatment mean on the data scale. This is a consequence of the link function for normally distributed data, $\eta_{ij} = \mu_{ij}$, called the *identity link*. Until Nelder and Wedderburn (1972) introduced GLMs, there was no reason for the data–model scale distinction to occur to anyone. Because statistical methods textbooks still focus on traditional ANOVA, data–model scale issues have yet to come to their attention.

What to Include in the Linear Predictor

For normally distributed data, the hard part of writing the model is getting the WWFD ANOVA right. The model equation follows. The linear predictor is defined by each line of the combined ANOVA except the last.

With non-normal data, writing the linear predictor requires more care. One must consider the last line of the combined ANOVA *and* the variance implied by the response variable's distribution. To illustrate, consider the randomized block example with normal data as opposed to the same design with binomial data. For normally distributed data, there is no ambiguity about the linear predictor. It consists of an intercept plus a term for each line in the ANOVA except the last, e.g., $\mu + \tau_i + b_j$ for the randomized block design. We used these terms to estimate the means, μ_{ij}. The last line of the ANOVA gives us mean square(residual), which we use as the estimate of the variance, σ^2.

What if we borrow the normal-data linear predictor and use it for binomial data? Using the linear predictor $\eta + \tau_i + b_j$ means not using any information from the last line of the ANOVA. This may be a problem. Why? It is reasonable to assume that the ijth experimental unit has unique characteristics not fully explained by the treatment and block effects alone. In the normal case, experimental unit uniqueness is implicit: e_{ij} appears in the model equation to account for it and σ^2 measures it. But, unlike the normal, the binomial does not have a separate variance term. Once you estimate p_{ij} you automatically have the mean and the variance. If the binomial model merely borrows the linear predictor from the normal model, experimental unit uniqueness is ignored. This results in the model accounting for less variation than is actually present in the data. This is one form of *overdispersion*, the modeling term for the observed variance exceeding the variance expected in theory. Overdispersion results in downward-biased standard errors, meaning that confidence intervals are too narrow, and upward-biased test statistics, meaning that Type I error rates are inflated, often severely.

The solution to this problem lies in accurately understanding the WWFD combined ANOVA. How this plays out depends on the specifics of the distribution. Examples below will show how this issue is dealt with in the two cases where the agronomist is most likely to see it—with binomial and count data.

Conditional and Marginal

Above, we explored the distinction between the conditional and marginal distribution of the data. In each of the three motivating examples, ANOVA on the untransformed data produced treatment mean estimates that were different from those produced by the corresponding GLMM. The binomial example referred to two different understandings of p_i, the conditional and marginal. While not explicitly mentioned in the count and continuous proportion examples, these distinctions were also present. The distinction between conditional and marginal inference exists for all non-normal data. Working effectively with non-normal data requires understanding the difference.

The conditional, or GLMM, estimate and the marginal estimate address two potentially useful, but distinctly different, questions. In the binomial case, the GLMM estimate says, "If I take an average member of the population—which means a member of the population whose block effect $b_j = 0$—what is the estimated binomial probability?" The marginal mean addresses the question, "If I average across all the members of the population, what is the mean proportion?" Because the marginal distribution is highly skewed, another way to think of this is to consider the mean and median. Because the distribution of block effects is symmetric, an estimate of the probability at $b_j = 0$ is closely related to the median. Household income provides a useful analogy because income data tend to be strongly right-skewed: median income accurately characterizes a typical household; mean income does not provide a useful characterization of a typical household, but it does accurately measure the amount of money in the overall economy. Conditional inference: median of a skewed distribution. Marginal inference: mean of a skewed distribution. Which is right? It depends. What is the question?

Because the normal distribution is symmetric, the conditional–marginal issue does not arise with LMs and LMMs. With GLMs and GLMMs, researchers must decide which understanding of "expected value" best addresses a study's objectives.

THREE EXAMPLES OF GENERALIZED LINEAR MIXED MODEL IMPLEMENTATION AND INTERPRETATION

The purpose of the three examples presented here is to illustrate the main issues researchers will encounter using GLMMs to analyze and interpret non-normally distributed data. This will necessarily be a survey of main issues; it cannot be exhaustive, nor can it go into great depth. Readers seeking greater breadth and detail are referred to texts such as Gbur et al. (2012), Stroup (2013a), and Faraday (2006). Additional references are noted as they arise for specialized applications suggested by these examples.

Specific software commands are avoided where possible. The SAS and R statements needed to implement the examples are available in the supplemental material. Each example concludes with a brief characterization of about how the GLMMs presented compare with common transformations with regard to Type I error control, power, and accuracy of estimates of the mean.

Example 1: Randomized Complete Block Design, Binomial Data

This example uses the same data as the binomial example discussed above. The design is a randomized complete block with eight blocks and two treatments. At each experimental unit, that is, at each block × treatment combination, 100 observations are made and the response of interest is the number out of those 100 that have a characteristic of interest. For example, how many seeds out of 100 germinated?

From the discussion above, the standard linear predictor for the RCBD is $\eta + \tau_i + b_j$. Because there were 100 observations per experimental unit, even if the probability of a seed germinating is 0.9, the minimum criterion for using the normal approximation to the binomial given in standard statistical methods textbooks, that

$N(1-p)$ exceed 5, or even 10, is easily satisfied. Figure 1 visually underscores this point. Hence the temptation to compute ANOVA with the sample proportion, $p_{ij} = y_{ij}/N$. However, we saw above the problems with that approach that were not apparent at the time standard statistical methods textbooks were written.

From the discussion above, the GLM for the RCBD requires

- the distribution of the observations: if block effects are considered random—strongly recommended, especially when working with non-normal data—then the conditional distribution of the observations given the blocks must be specified: $y_{ij}|b_j \sim$ Binomial(N,p_{ij}), where $N = 100$ is the number of observations per experimental unit and p_{ij} is the probability of the outcome of interest (e.g., seed germinates) for the ith treatment and jth block
- the distribution of block effects if they are considered random: the standard assumption is that block effects are normally and independently distributed with mean 0 and variance σ_B^2; in statistical notation, $b_j \sim NI(0,\sigma_B^2)$.
- the link function: $\eta_{ij} = \text{logit}(p_{ij}) = \log[p_{ij}/(1-p_{ij})]$
- the linear predictor: $\eta_{ij} = \eta + \tau_i + b_j$

When implemented, this yields an F value for treatment of 68.76 with a p value <0.0001. Recall that the p values from the example discussed above ranged from 0.0355 to 0.1132. This discrepancy is an example of overdispersion-induced test-statistic inflation. From the discussion concerning what to include in the linear predictor, the model that should be fitted is a modified GLMM adding an effect corresponding to the "unit(block)|treatment" line of the WWFD ANOVA. The amended model:

- linear predictor: $\eta_{ij} = \eta + \tau_i + b_j + u_{ij}$, where u_{ij} is the random unit-level effect; with blocked designs, u_{ij} is mathematically equivalent to the block × treatment interaction; the linear predictor can be written equivalently as $\eta_{ij} = \eta + \tau_i + b_j + (bt)_{ij}$
- random effect distribution: whether referred to as the unit-level effect or block-treatment effect, its distribution is assumed to be normal, independent, with mean 0 and variance σ_U^2 (or, if you prefer, σ_{BT}^2)
- distribution of the observations: $y_{ij}|b_j,u_{ij} \sim$ Binomial(N,p_{ij}), which is simply the conditional distribution from above with the unit-level random effect added

This yields the p value 0.0355 and the estimates of the probabilities in Table 2 for the GLMM. In addition to overdispersion, this example also illustrates the hazard of "borrowing" the linear predictor intact from standard analysis and using it for non-normal data. Users must be aware of the variance of their response variable's distribution in conjunction with the interpretation of the last line of the WWFD ANOVA.

We noted above that there are two ways of understanding treatment means with non-normal data: the conditional and the marginal. In this example, the conditional mean is the ith treatment probability, p_i, for an average member of the population; the marginal mean is the average of the probabilities across all members of the population. Use the conditional mean if the research objectives focus on what would be expected for a typical producer, typical farm, etc. The above GLMM gives these estimates, as well as accurate test statistics, standard errors, confidence intervals, etc.

However, what if the marginal mean is more appropriate for a given research objective? We saw above that the normal approximation of the binomial, implemented with standard ANOVA, yields unbiased estimates of the marginal means but not the correct standard errors. As a consequence, ANOVA produces inaccurate confidence intervals and invalid test statistics. Transformations do not help. Stroup (2013b) showed that, if anything, the angular transformation is even more inaccurate than the untransformed ANOVA—and it lacks power to test for treatment differences. For binomial data, ANOVA with or without transformation should be considered unacceptable for scientific publication. If the marginal mean best addresses the research objectives, the correct approach requires an alternative formulation of the GLMM.

The alternative GLMM needed for marginal means begins with the GLMM we just considered, whose linear predictor was $\eta_{ij} = \eta + \tau_i + b_j + u_{ij}$ with random block and unit-level effects. The variance of the linear predictor is $\sigma_B^2 + \sigma_U^2$ and the correlation between the pair of unit effects within the same block is $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_U^2)$. In the early days of GLMs, when computers could handle GLMs but not GLMMs, these variance and covariance results provided an insight that allowed GLM software to handle some GLMMs. Zeger et al. (1988) presented a method that removed the random effects from the linear predictor, making the model a GLM instead of a GLMM. To account for the unit and block effects, they created a structure they called a *working correlation*. Instead of requiring the entire distribution to be specified, they showed that it was sufficient merely to identify the mean, the variance, and the correlation structure to implement a GLM analysis. Instead of a distribution and likelihood, as in our brief excursion into probability and estimation theory above, they now had what statistical theorists call a *quasi-likelihood*. Analysis with quasi-likelihood and a working correlation can be computed using what Zeger et al. (1988) called *generalized estimating equations* (GEEs). Models defined this way are referred to as *GEE models*. Both the R package GEE and SAS PROC GLIMMIX can implement GEE models.

The GEE provides valid analysis for marginal means. Because it focuses on the marginal mean, it is also referred to as a *marginal model*. For this example, its elements are:

- linear predictor: $\eta_{ij} = \eta + \tau_i$
- quasi-likelihood (replaces distribution): y_{ij} has a Binomial(N,p) quasi-likelihood with variance $\phi_W p_{ij}(1-p_{ij})$ and covariance $\phi_W \rho_W [p_{0j} p_{1j} (1-p_{0j})(1-p_{1j})]^{1/2}$, where the W subscripts in ϕ_W and ρ_W denote "working"; ϕ_W assumes the role that $\sigma_B^2 + \sigma_U^2$ plays in the GLMM, while ρ_W replaces the interclass correlation $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_U^2)$
- link function: $\eta_{ij} = \text{logit}(p_{ij}) = \log[p_{ij}/(1-p_{ij})]$

Another model that targets the marginal mean is a GLMM using the beta distribution. This approach is technically valid only when all experimental units have the same number of yes–no observations, that is when $N_{ij} = N$. If N_{ij} vary but are more or less equal, assuming a beta distribution gives acceptable results. The model is:

- distribution: $y_{ij}|b_i \sim$ Binomial(N,p_{ij})
- $p_{ij} \sim$ Beta(π_{ij},φ)
- link function: logit(p_{ij})
- use response variable $\text{pct}_{ij} = y_{ij}/N$ (software note: using pct is not the same as specifying binomial y/N; SAS and R have different conventions for this and it is important to get them right)

Generalized estimating equation analysis of the data in this example yields a p value of 0.1132 and estimates of the treatment probabilities of 0.90 ± 0.05 and 0.74 ± 0.07 for the control and test treatments, respectively. The beta model yields a p value of 0.0713 and estimated treatment probabilities of 0.88 ± 0.04 and 0.74 ± 0.06. The GEE p value and probability estimates are identical to the standard ANOVA using the normal approximation, but, unlike ANOVA, the GEE standard errors reflect the mean–variance relationship of the binomial. The beta shows a lower p value and slightly different treatment mean estimates.

Stroup (2013b) found that the beta GLMM yielded the most robust combination of Type I error control, power for detecting treatment differences, and accurate confidence interval coverage. For the GEE, power was up to 10% lower and confidence interval coverage was less accurate. Assuming equal N for all experimental units, the beta GLMM is the preferred method if the marginal mean is the appropriate target. For unequal N, use the GEE. Either way, these are the appropriate ways to analyze binomial data when the marginal mean is the target, that is, when the research question is, "If I average across all the members of the population, what is the mean proportion and how do these proportions differ by treatment?"

Example 2: Split-Plot Experiment, Count Data

The split plot is arguably the most common design structure in plant and soil science research. Such experiments involve two or more treatment factors. Typically, large units called *whole plots* are grouped in blocks. Levels of one factor, called the *whole-plot factor*, are randomly assigned to whole plots. Each whole plot is divided into smaller units, called *split plots*. Levels of the second factor are randomly assigned to split-plot units within each whole plot. In this example, there are six blocks, a whole-plot factor with two levels, referred to here as A1 and A2, and a split-plot factor with two levels, referred to as B1 and B2. While the ANOVA structure is undoubtedly familiar to readers, reviewing the WWFD ANOVA process makes it easier to understand modeling options for non-normal data and to distinguish those that make sense from those that do not.

Recall that WWFD involves describing the topographical structure (experiment design), the treatment structure, and then combining them. In the split plot, the experiment design consists of blocks, whole-plot units, and split-plot units; the treatment design is composed of Factor A (whole plot) and Factor B (split plot). Table 6 summarizes the WWFD process.

The textbook model equation for this ANOVA is $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + r_k + w_{ik} + s_{ijk}$, where α and β refer to treatment (Factors A and B) effects, r refers to block effects (r is used here instead of b to avoid confusion with β), w refers to whole-plot effects, and s refers to split-plot effects. Translating the model equation to a linear predictor yields $\eta_{ijk} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + r_k + w_{ik}$. Assuming normally distributed data means assuming that the observations, given the design structure, have a normal distribution—in statistical notation, $y_{ijk}|r_k,w_{ik} \sim NI(\mu_{ijk},\sigma_S^2)$. Because observations are taken at the split-plot level, conditional on the design effects, they have variance associated with the split plot. Block and whole-plot effects are also assumed to contribute variation: $r_k \sim NI(0,\sigma_R^2)$ and $w_{ik} \sim NI(0,\sigma_W^2)$. The model uses the linear predictor, η_{ijk}, to estimate the means of the observations, μ_{ijk}, and all inference follows from there.

All of this is familiar to any plant or soil scientist who has analyzed a split plot. How literally does it adapt to count data? Classical probability theory assumes that counts have a Poisson distribution. Borrowing the linear predictor from the ANOVA-based model equation and using the canonical parameter as the link function gives a preliminary GLMM for the split-plot experiment with count data:

- distribution: $y_{ijk}|r_k,w_{ik} \sim \text{Poisson}(\lambda_{ijk})$
- link function: $\eta_{ijk} = \log(\lambda_{ijk})$
- linear predictor: $\eta_{ijk} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + r_k + w_{ik}$
- distribution of block and whole-plot effects: as given above

I refer to this as the *naive Poisson model*. Recall that this is exactly the strategy used to construct the preliminary model for the randomized block with binomial data in Example 1, and it proved to be inadequate. For the same reasons, this model is likely to be inadequate: like the binomial, the Poisson is a one-parameter distribution and hence the model needs some provision for accounting for unit-level—in this case the split-plot unit level—variation. In the binomial case, the last line of the ANOVA was restored to the linear predictor.

One could do the same thing here. The revised model would use the same distribution and link, but the linear predictor would be $\eta_{ijk} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + r_k + w_{ik} + s_{ijk}$. An alternative—and preferable—approach is to leave the linear predictor intact but change the assumed distribution to negative binomial, that is $y_{ijk}|r_k,w_{ik} \sim \text{NB}(\lambda_{ijk},\phi)$. Assuming a negative binomial actually includes a split-plot unit effect but with a different distribution.

Aside: For those with a probability background, a helpful way to think of the negative binomial model is as follows. Assume the conditional distribution of the observations given the random block, whole-plot, and split-plot effects is $y_{ijk}|r_k,w_{ik},u_{ijk} \sim \text{Poisson}(\lambda_{ijk}u_{ijk})$, where $u_{ijk} \sim \text{Gamma}(1/\phi,\phi)$. The resulting distribution of $y_{ijk}|r_k,w_{ik}$ is $\text{NB}(\lambda_{ijk},\phi)$. The link function is

Table 6. "What Would Fisher Do" ANOVAs for a split-plot experiment with a blocked whole plot.

Topographical		Treatment		Combined	
Source	df	Source	df	Source	df
Block	5			Block	5
		A	1	A	1
Whole plot(block)	6			Whole plot|A (whole plot error or block × A)	6 – 1 = 5
		B	1	B	1
		A × B	1	A × B	1
Split-plot unit (whole plot)	12	"Parallels"	20	Split-plot unit|B ("residual" or block × A × B)	12 – 2 = 10
Total	23	Total	23	Total	23

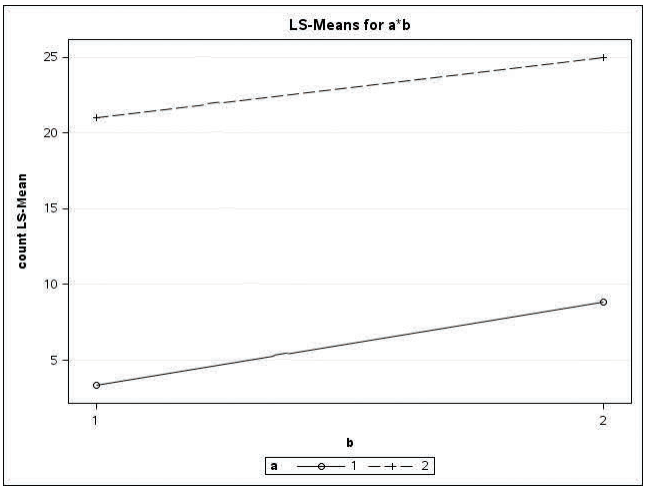

Fig. 3. Plot of A × B treatment combination least squares (LS) means for count data.

$\log(\lambda_{ijk}u_{ijk}) = \log(\lambda_{ijk}) + \log(u_{ijk})$. In other words, $\log(u_{ijk})$ in the negative binomial model replaces s_{ijk} in the Poisson model. Both account for split-plot-level effects.

Why the negative binomial model and not the Poisson? The probability theory underlying the negative binomial is consistent with biological theory of how counts develop in field situations—at least more so than the theory underlying the Poisson. Moreover, studies with biological counts generally confirm that the negative binomial does provide the more accurate characterization of the observed variation.

Why include all this seemingly mathematical minutiae here? Because it illustrates that the ANOVA structure Fisher devised remains as relevant as ever as a basis for analyzing experimental data, but slavishly attaching ANOVA recipes tied to normal theory when the data are clearly non-normal will not do. The ANOVA tells us what sources of variation have probability distributions, but it does *not* say that those distributions have to be normal. Clear thinking about the experimental context is essential.

Figure 3 shows a plot of the treatment combination means. The plot suggests negligible A × B interaction, negligible main effect of B, but a noticeable main effect of A. Obviously, whether these effects are statistically significant or not depends on inferential statistics from a valid statistical model.

The naive Poisson model with linear predictor borrowed from textbook split-plot ANOVA yields the following

Source	Num DF	Den DF	*F* value	*p*-value
a	1	5	11.40	0.0198
b	1	10	15.81	0.0026
a × b	1	10	7.67	0.0198

Unexpectedly, all effects would be declared highly significant. This illustrates the problem with the naive Poisson model: it does not use information from the last line of the WWFD combined ANOVA and as a result shows symptoms of overdispersion. This is confirmed by an overdispersion diagnostic statistic available with PROC GLIMMIX when run using either Laplace or quadrature: the Pearson $\chi^2/df = 7.4$. In theory, a model with no overdispersion will produce a Pearson χ^2/df of approximately 1. In practice, this statistic should not appreciably exceed 1. A value over 2 should be considered evidence of possible overdispersion; 7.4 is decisive evidence.

An alternative diagnostic, available with R's LME4 package as well as SAS GLIMMIX uses the Akaike information criterion (AIC) fit criterion. The naive Poisson model above yields an AIC of 358 compared with the negative binomial model AIC of 178. The smaller AIC implies the better model; in this case, it is not subtle.

The negative binomial model yields the following results for the tests of treatment effects:

Source	Num df	Den df	*F* value	*p* value
a	1	5	8.59	0.0326
b	1	10	1.74	0.2168
a × b	1	10	0.77	0.4005

The results are distinctly more consistent with the mean plot in Fig. 3.

In the era between 1990 and 2005, good GLM software (e.g., SAS PROC GENMOD) was available, but good GLMM software had not yet appeared. During this interregnum, GEE was often used to analyze split-plot experiments with non-normal data. The "tradition" persists in certain disciplines associated with plant and soil science. Recall from the binomial Example 1 that the GEE replaces random effects in the linear predictor with working variance and correlation and replaces the distribution with a quasi-likelihood. The GEE model for this example is

- quasi-likelihood of y_{ijk}: Poisson(λ_{ijk}) quasi-likelihood with variance $\phi_W\lambda_{ijk}$ and covariance $\phi_W\rho_W\sqrt{(\lambda_{1jk}\lambda_{2jk})}$
- link function: $\eta_{ijk} = \log(\lambda_{ijk})$
- linear predictor: $\eta_{ijk} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + r_k$; because the GEE does not accommodate random effects, the block effect, r_k, must be assumed to be fixed

The GEE yields the following treatment effect tests:

Source	Num df	Den df	*F* value	*p* value
a	1	5	13.77	0.0138
b	1	10	0.81	0.3888
a × b	1	10	0.39	0.5444

For this example, GEE leads to conclusions similar to those that follow from the negative binomial model. However, Stroup (2013b) found that the negative binomial GLMM yields robust performance in terms of Type I error control, power, and confidence interval coverage. In contrast, GEE models showed >15% power loss and erratic confidence interval coverage. Variations of Poisson-based GLMMs did no better, their worst problem being erratic Type I error control and downward-biased estimates of the mean.

To illustrate the confidence interval coverage issue, Table 7 shows the estimates of the mean rates from the naive Poisson model shown above, the negative binomial GLMM, and GEE. These results are typical of what Stroup (2013b) found.

The sample means are marginal means, so they are shifted upward relative to the true mean count. This partly accounts for the poor coverage using standard ANOVA with untransformed data. The Poisson GLMM and GEE overcompensate, shifting estimates too far downward. Only the negative binomial model obtains consistently accurate estimates. Stroup (2013b) also obtained results for transformations–notably the logarithmic and square root–commonly recommended for count data. Estimation accuracy was erratic and power loss was considerable (as high as 40%). As with binomial data, transformations in a mixed model setting not only do not help, they tend to make things worse.

Table 7. Estimated mean counts produced by different models.

A	B	Sample mean	Poisson generalized linear mixed model	Negative binomial generalized linear mixed model	Generalized estimating equation
1	1	3.3	2.7	3.2	2.4
1	2	8.8	7.1	8.6	6.5
2	1	21.0	14.2	19.1	15.4
2	2	25.0	16.9	23.3	18.3

Example 3: Repeated Measures, Binomial Data

Repeated measures occur when observations are taken at planned times, e.g., at defined growth stages or at regular intervals during the growing season, on the same experimental unit. Repeated measures also occur in space, for example when measurements are taken at regularly spaced depths from a soil core sample. Whether in space or time, the defining modeling consideration is correlation among the measurements on the same experimental unit. Observations on the same experimental unit are not independent and their correlation depends on distance: adjacent observations tend to be more highly correlated than observations farther apart.

Much has been written on repeated measures analysis for normally distributed data. Readers are referred to Littell et al. (2006), Gbur et al. (2012), and Stroup (2013a) for comprehensive introductions to repeated measures LMMs. While basic LMM principles of modeling correlated data apply to non-normal data, these similarities may not be readily apparent. As with the other models considered here, non-normality introduces unique considerations that would not occur in a normal-data-only world. Gbur et al. (2012) and Stroup (2013a) covered this topic in detail. Here, the essentials and the bottom line are introduced.

The two primary modeling approaches are those that use a working correlation—similar to the GEE models shown in Examples 1 and 2—and those that embed a correlated unit-level effect in the linear predictor, similar to the binomial with linear predictor $\eta_{ij} = \eta + \tau_i + b_j + u_{ij}$ from Example 1. The GEE approach dates from its introduction by Zeger et al. (1988). The GLM software introduced in the 1990s could implement GEEs. On the other hand, true repeated measures GLMMs could not be fully implemented until the 2008 release of SAS PROC GLIMMIX. At this time, there still is no equivalent capability in R. For this reason, the repeated measures GLMM, despite its advantages, is much less well known. As an aid to readers negotiating supplemental reading, Gbur et al. (2012), Stroup (2013a), and SAS GLIMMIX documentation refer to the GEE as *R-side modeling* and the true GLMM as *G-side modeling*.

This example has two treatments, generically labeled 0 and 1, and 10 experimental units (called *plots*) randomly assigned to each treatment in a completely randomized design. Measurements are taken on each experimental unit at five times. For example, the two treatments could be different management practices and the times could be 4, 8, 12, 16, and 20 wk after mowing. Table 8 shows the WWFD ANOVA for this design.

Superficially, the combined ANOVA looks like a split-plot ANOVA. Indeed, one frequently used model for agronomic repeated measures data is called the "split plot in time." This model is equivalent to assuming compound symmetry—that is, all repeated measures are equally correlated regardless of how far apart they are in space or time. This is sometimes—but certainly not always—true of agronomic data. Failure to account for distance-dependent correlation is a form of overdispersion, with the same consequences seen in the previous examples.

In this example, the data observed are binomial, with 50 yes–no observations per experimental unit per measurement occasion. The GLMM that follows from the WWFD combined ANOVA is

- linear predictor: $\eta_{ijk} = \eta + \alpha_i + \tau_k + (\alpha\tau)_{ik} + r(\alpha)_{ij} + w_{ijk}$, where α denotes treatment, τ denotes time, r denotes plot, and w denotes within-plot measurement occasion
- random effect distributions: plot, $r(\alpha)_{ij}$, more commonly denoted b_{ij} (for "between subject" effect) $\sim NI(0,\sigma_B^2)$; the five within-subject effects on each plot, denoted $[w_{ij1}\ w_{ij2}\ w_{ij3}\ w_{ij4}\ w_{ij5}]$ are assumed to have a multivariate normal distribution, that is, each w_{ijk} has a normal distribution and each pair is correlated; the mean of each effect is 0, but the form of the correlation structure can vary (see below)
- response distribution: $y_{ijk}|r(\alpha)_{ij}, w_{ijk} \sim \text{Binomial}(N, p_{ijk})$ where i references treatment, j references plot, k references time, and $N = 50$
- link function: $\eta_{ijk} = \text{logit}(p_{ijk})$

Gbur et al. (2012) and Stroup (2013a) presented common correlation structures, with Gbur et al. focusing on those most relevant to plant and soil science research. The structure that best fits these data, using the model selection procedure described below, is the first-order autoregressive [AR(1)] model with variance σ_W^2 and the correlation ρ^d where d is the distance between two observations. For example, the distance between the first and second repeated measures, w_{ij1} and w_{ij2}, is 1; the distance between the first and third, w_{ij1} and w_{ij3}, is 2; and so

Table 8. "What Would Fisher Do" ANOVA for repeated measures example.

Topographical		Treatment		Combined	
Source	df	Source	df	Source	df
		Treatment	1	Treatment	1
Plots	19			Plots|treatment (between subjects)	19 – 1 = 18
		Time	4	Time	4
		Time × treatment	4	Time × treatment	4
Measurement occasion (plot)	80	"Parallels"	90	Occasion (plot)|time (within subjects)	80 – 8 = 72
Total	99	Total	99	Total	99

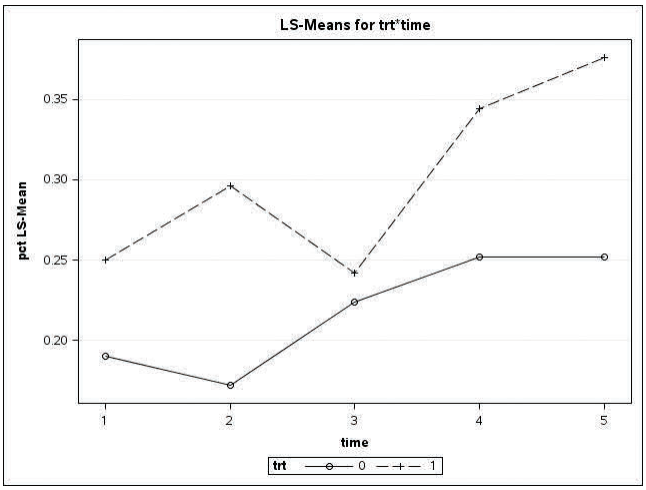

Fig. 4. Plot of sample proportions (inverse linked yes/no least squares [LS] mean) with time by treatment for Example 3.

forth. The correlation structure is most succinctly described in matrix form, which for the AR(1) is

$$\mathrm{Var}\begin{bmatrix} w_{ij1} \\ w_{ij2} \\ w_{ij3} \\ w_{ij4} \\ w_{ij5} \end{bmatrix} = \sigma_W^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$$

The diagonal elements give the variance. The off-diagonal elements give the correlation between times cross-referencing row and column; for example, the first row and third column gives the correlation between the first and third repeated measurement.

Analysis of the repeated measures GLMM begins by determining which of the plausible covariance structures best fits the data. Typically, this is done by computing fit criteria, e.g., the small-sample corrected Akaike information criterion (AICC). Do this by fitting each competing covariance structure in SAS GLIMMIX using the METHOD = LAPLACE option. At this time, this step must be done with SAS GLIMMIX because it is currently the only GLMM software that supports modeling G-side correlated effects. For these data, the AICC for the split-plot-in-time model (i.e., assuming compound symmetry) is 646 and for the AR(1) is 626. Smaller is better, meaning that there is non-negligible correlation that must be accounted for in subsequent analysis. While not shown here, in practice other covariance structures—in particular the first-order ante-dependence [ANTE(1)] model—should be considered. In my experience, the vast majority of data fit the split-plot-in-time, AR(1), or ANTE(1) structure. The unstructured model is similar to the multivariate analysis of variance (MANOVA), a procedure that was often used in pre-LMM days. With LMMs and GLMMs available, using MANOVA for repeated measures is extreme overkill, inefficient and underpowered. With non-normal data it is also prone to computational difficulties. The bottom line on unstructured covariance with non-normal data: needless effort prone to intractable computational problems, yielding inferior inferential statistics even on the occasions when such statistics can be obtained. Don't bother.

Figure 4 shows the mean plot for these data.
The AR(1) GLMM tests for the treatment and time effects are

Source	Num df	Den df	*F* value	*p* value
trt	1	9	4.90	0.0542
time	4	57.99	2.82	0.0330
trt × time	4	57.98	1.74	0.1536

Notice the non-integer values of the denominator df. They, and the *F* and *p* values, reflect the procedure developed by Kenward and Roger (2009) to account for the effect of the covariance structure on degrees of freedom and standard errors. Although the Kenward–Roger adjustment was derived for the LMM with normally distributed data and is an ad hoc procedure for GLMMs with non-normal data, informal simulation studies consistently have suggested that the adjustment is accurate. The Kenward–Roger adjustment requires that the SAS GLIMMIX default computing algorithm, pseudo-likelihood, be used rather than the Laplace algorithm used to obtain AICC statistics. Stroup (2013b) found that for binomial and Poisson GLMMs, pseudo-likelihood with the Kenward–Roger adjustment yields better Type I error control than Laplace while preserving the GLMM's advantage with respect to power and accuracy in estimating treatment means.

While technically nonsignificant, the test of treatment × time interaction should not be ignored. In the first place, it is a multiple-degrees-of-freedom test with a *p* value <0.20. Standard statistics methods textbooks advise investigating simple effects for such cases. Inspection of Fig. 4 suggests that treatment differences vary noticeably with time. There are many ways to address this. The most appropriate depend on the particulars of the research and its objectives. Two that are shown here are "slices" by treatment, testing whether there is a change in response with time on a treatment-by-treatment basis, and simple-effect tests of treatment effect for each time. These are shown below, along with the results for the GEE, to allow a side-by-side comparison.

The primary alternative is GEE. The disadvantages of GEE are: (i) no comparison of covariance structures is possible (the AICC and related fit statistics are undefined for GEEs); and (ii) assuming the GLMM is properly specified, its power and treatment mean estimation accuracy are greater than those of GEE with no sacrifice in Type I error control. The advantage of GEE: theory suggests that it may be more robust to misspecified models, meaning that while it may never be *exactly* correct it may be more likely to be *approximately* correct than the GLMM.

For these data, the GEE model is

- linear predictor: $\eta_{ijk} = \eta + \alpha_i + \tau_k + (\alpha\tau)_{ik} + r(\alpha)_{ij}$, where α denotes treatment, τ denotes time, and r denotes plot
- random effect distributions: none—the GEE is a strictly fixed-effects model
- response distribution of y_{ijk}: Binomial(N,p_{ijk}) quasi-likelihood, where i references treatment, j references plot, k references time, and $N = 50$
- link function: $\eta_{ijk} = \mathrm{logit}(p_{ijk})$
- working covariance:

$$\text{working Var}\left(y_{ijk}\right) = \phi_W p_{ijk}\left(1 - p_{ijk}\right)$$

Table 9. Treatment mean estimates from generalized estimating equation (GEE) and generalized linear mixed model (GLMM) repeated measures analyses.

Observation	Treatment	Time	P_hat_gee†	se_P_gee‡	P_hat_glmm§	se_P _glmm¶
1	0	1	0.17921	0.030499	0.18006	0.035363
2	0	2	0.16141	0.029131	0.16314	0.032964
3	0	3	0.21293	0.032756	0.21098	0.039351
4	0	4	0.24162	0.034407	0.23686	0.042445
5	0	5	0.24175	0.034409	0.23800	0.042603
6	1	1	0.24021	0.034314	0.23380	0.042155
7	1	2	0.28705	0.036556	0.28741	0.047381
8	1	3	0.23047	0.033746	0.22583	0.041181
9	1	4	0.33693	0.038348	0.32785	0.051061
10	1	5	0.37100	0.039279	0.36654	0.053284

† P_hat_GEE denotes estimated $\hat{P}_{ij}$ for each treatment–time combination from GEE.

‡ se_P_GEE denotes associated standard error.

§ P_hat_GLMM denotes estimated $\hat{P}_{ij}$ for each treatment–time combination from GLMM.

¶ se_P_GLMM denotes associated standard error.

Table 10. Simple effect estimates from generalized linear mixed model (GLMM)-based repeated measures analysis.

Tests of effect slices for treatment × time
Sliced by treatment (trt)

trt	Numerator df	Denominator df	*F* value	Pr > *F*
0	4	62.74	1.20	0.3186
1	4	53.17	3.40	0.0151

Simple effect level	trt	_trt	Odds ratio	*p* value
Time 1	0	1	0.720	0.2562
Time 2	0	1	0.483	0.0151
Time 3	0	1	0.917	0.7598
Time 4	0	1	0.636	0.1141
Time 5	0	1	0.540	0.0334

$$\text{working Cov}\left[y_{ijk}, y_{ijk'}\right] = \begin{cases} \phi_W \rho_{W_{k,k'}} \sqrt{p_{ijk}\left(1-p_{ijk}\right)}\sqrt{p_{ijk'}\left(1-p_{ijk'}\right)} \\ \qquad\qquad \text{if } ij \text{ same for both } y \\ 0 \text{ otherwise} \end{cases}$$

The terms ϕ_W and $\rho_{Wk,k'}$ denote the working scale and correlation, respectively. The form of the working covariance can vary in the same manner as the covariance of unit-level effects does in GLMMs. For example, the AR(1) working covariance is

$$\text{Cov}_W \begin{bmatrix} y_{ij1} \\ y_{ij2} \\ y_{ij3} \\ y_{ij4} \\ y_{ij5} \end{bmatrix} = \phi_W \begin{bmatrix} 1 & \rho_W & \rho_W^2 & \rho_W^3 & \rho_W^4 \\ & 1 & \rho_W & \rho_W^2 & \rho_W^3 \\ & & 1 & \rho_W & \rho_W^2 \\ & & & 1 & \rho_W \\ & & & & 1 \end{bmatrix}$$

The AR(1) GEE tests for the treatment and time effects are

Source	Num df	Den df	*F* value	*p* value
trt	1	14.63	7.69	0.0145
time	4	64.84	2.70	0.0382
trt × time	4	64.85	1.55	0.1985

Notice that the *p* value for the treatment × time effect is just under 0.20 for the GEE compared with just over 0.15 for the GLMM. This is where the power loss of the GEE tends to appear in repeated measures analyses.

Table 9 shows the estimated probabilities for the GLMM and GEE analyses. Tables 10 and 11 show simple effects—slices by treatment and treatment differences by time—from the GLMM and GEE analyses, respectively.

The most striking difference between the two analyses is the lack of difference. The probabilities and associated standard errors are close, with differences appearing mostly in the third decimal place. The GEE means show a slight shift toward 0.5, expected because they are estimates of the marginal mean, not true probability estimates. The slice and simple-effect results would lead to essentially identical conclusions: there is evidence of changes with time under Treatment 1 but not under Treatment 0 and there is evidence of treatment difference at Times 2 and 5. The one discrepancy appears at Treatment 4: the GEE shows a *p* value of 0.07 compared with 0.11 for the GLMM. Depending how rigidly a policy regarding α is imposed, this could create an issue. It shouldn't, because the estimated odd ratios are nearly identical. This is an illustration of why journals would be well advised to tone down emphasis on *p* values and pay more attention to interval estimates.

SUMMARY AND CONCLUSIONS

This discussion has attempted to give plant and soil science researchers, and the statistical scientists with whom they collaborate, a sense of the issues and methods associated with contemporary analysis of non-normal data. To borrow a phrase from a recent advertising campaign, "This is not your father's

Table 11. Simple effect estimates from generalized estimating equation (GEE)-based repeated measures analysis.

Tests of effect slices for treatment × time Sliced by treatment (trt)				
trt	Numerator df	Denominator df	*F* value	Pr > *F*
0	4	64.8	1.08	0.3716
1	4	64.78	3.21	0.0182

Simple effect level	trt	_trt	Odds ratio	*p* value
Time 1	0	1	0.691	0.1900
Time 2	0	1	0.478	0.0109
Time 3	0	1	0.903	0.7095
Time 4	0	1	0.627	0.0720
Time 5	0	1	0.541	0.0183

statistical analysis." If I have left the impression that the GLMMs have a steep learning curve, that is right. The question, then, is: given that modern researchers have a myriad of new concerns to deal with, is GLMM just another example of the statistical tail wagging the scientific research dog, or does the GLMM really matter despite the learning curve?

To answer this question, here are my take-home messages:

- Standard ANOVA on untransformed non-normal data depends on the Central Limit Theorem, which says that treatment means have an approximate normal distribution if the sample size is *large enough*.
- In a world where four replications are typical and budgets are tight, *large enough* is problematic.
- Even if normality holds, homogeneity of variance does not, except in the usually uninteresting case of no difference among treatment means.
- As a consequence, standard ANOVA on untransformed non-normal data suffers loss of power—often *severe* loss of power—and inaccurate estimates of treatment means.
- Research budgets being what they are—and what they will probably be for the foreseeable future—lavishly replicated experiments are not abundant in plant and soil science research. On the other hand, minimally replicated, often underpowered experiments are common. It makes no sense to further handicap such experiments with the power loss associated with analysis using standard ANOVA.
- Transformations not only do not help, they are counterproductive. The theory underlying transformations was developed in a world where mixed models did not exist. There is mounting evidence that transformations do more harm than good for the models required by the vast majority of contemporary plant and soil science research. This includes designs with any kind of blocking, from simple RCBDs through split-plot, repeated measures, and certainly any more complex designs. Reviewers need to cease and desist from suggesting transformations for non-normal data.
- Some suggest using non-parametric statistics. While non-parametric statistics do not assume normality, they are focused only on testing, not on estimation. In most plant and soil science research, the question is not, "Is there a treatment difference?" Instead, it is, "We know there is a difference. How big is it?" Non-parametric statistics are useless for the latter.
- Generalized linear mixed models provide statistically sound ways to address these issues. Small-sample investigations are providing an increasing body of evidence that GLMMs work as well in practice as they do in theory. The difference between now and before 2005 is that while the theory was incubating before 2005, texts to provide guidance and good, useable software did not exist, whereas now both are readily available.

In other words, for non-normal data, ANOVA, with or without transformed data, won't do. The loss of accuracy and power are too great. Given the current state of the art and the resources available to plant and soil science researchers, GLMMs and, in certain cases, GEEs are the methods of choice.

Admittedly, the learning curve is steep. Admittedly, a poorly chosen GLMM, or even a well-chosen but ineptly implemented GLMM, may be worse than standard ANOVA. Data analysts must pay attention to issues such as marginal vs. conditional models and inference, GEE vs. pseudo-likelihood vs. Laplace vs. quadrature, choices among distributions and particulars about those distributions for given types of response variables—topics that never occurred in traditional ANOVA and regression.

However, contemporary statistical practice, and eventually contemporary statistical training for future researchers, must adjust. Standard statistical practice as it was understood for most of the 20th century was a dramatic advance. But it was also a product of its times—the 1920s through 1940s—when the "computer" was a pencil and paper or at best a mechanical calculator and GLMM theory was 50 yr in the future. Nothing in science can remain frozen in time.

In the novel *Arrowsmith*, by Sinclair Lewis (1925), winner of the 1926 Pulitzer Prize, the protagonist, Martin Arrowsmith, is an up-and-coming medical researcher. While his primary loves are biology and biochemistry, his mentor, Max Gottlieb, implores his protégé to master the mathematics, even when the effort seems beyond him. Responsible, high-quality research, Gottlieb tells his student, is not possible unless it is quantitatively rigorous. While we may prefer Lewis to have refined Max Gottlieb's word choice, replacing "mathematics" with "statistics," his advice to the young Arrowsmith rings as true today as it did when the novel was written in 1925.

REFERENCES

Bartlett, M.S. 1947. The use of transformations. Biometrics 3:39–52. doi:10.2307/3001536

Breslow, N.E. and D.G. Clayton. 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88:9–25.

Eisenhart, C. 1947. The assumptions underlying analysis of variance. Biometrics 3:1–21. doi:10.2307/3001534

Faraday, J.J. 2006. Extending the linear model with R. CRC Press, Boca Raton, FL.

Federer, W.T. 1955. Experimental design. MacMillan, New York.

Fisher, R.A. 1925. Statistical methods for research workers. Oliver and Boyd, Edinburgh, UK.

Fisher, R.A. 1935. The design of experiments. Oliver and Boyd, Edinburgh, UK.

Fisher, R.A., and W.A. Mackenzie. 1923. Studies in crop variation: II. The manurial response of different potato varieties. J. Agric. Sci. 13:311–320. doi:10.1017/S0021859600003592

Gbur, E.E., W.W. Stroup, K.S. McCarter, S. Durham, L.J. Young, M. Christman, et al. 2012. Analysis of generalized linear mixed models in the agricultural and natural resources sciences. ASA, CSSA, and SSSA, Madison, WI.

Graybill, F.A. 1976. Theory and application of the linear model. Duxbury Press, North Scituate, MA.

Harville, D.A. 1976. Extensions of the Gauss–Markov theorem to include the estimation of random effects. Ann. Stat. 4:384–395. doi:10.1214/aos/1176343414

Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72:320–338. doi:10.1080/01621459.1977.10480998

Henderson, C.R. 1953. Estimation of variance and covariance components. Biometrics 9:226–252. doi:10.2307/3001853

Henderson, C.R. 1963. Selection index and expected genetic advance. In: W.D. Hanson and H.F. Robinson, editors, Statistical genetics and plant breeding. Natl. Res. Council Publ. 982. Natl. Acad. Sci., Washington, DC. p. 141–163.

Kenward, M.G., and J.H. Roger. 2009. An improved approximation to the precision of fixed effects from restricted maximum likelihood. Comput. Stat. Data Anal. 53:2583–2595. doi:10.1016/j.csda.2008.12.013

Kuehl, R.O. 2000. Design of experiments: Statistical principles of research design and analysis. 2nd ed. Duxbury Press, Pacific Grove, CA.

Laird, N.M., and J.H. Ware. 1982. Random-effects models for longitudinal data. Biometrics 38:963–973. doi:10.2307/2529876

Lewis, S. 1925. Arrowsmith. Harcourt, Brace & Co., San Diego.

Liang, K.-Y., and S.L. Zeger. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22. doi:10.1093/biomet/73.1.13

Littell, R.C., G.A. Milliken, R.D. Wolfinger, W.W. Stroup, and O. Schabenberger. 2006. SAS for mixed models. 2nd ed. SAS Inst., Cary, NC.

McCullagh, P., and J.A. Nelder. 1989. Generalized linear models. 2nd ed. Chapman and Hall, London.

Miller, R.G. 1997. Beyond ANOVA: Basics of applied statistics. CRC Press, Boca Raton, FL.

Milliken, G.A., and D.E. Johnson. 2009. Analysis of messy data. Vol. 1. 2nd ed. Chapman and Hall, New York.

Nelder, J.A., and R.W.M. Wedderburn. 1972. Generalized linear models. J. R. Stat. Soc. A 135:370–384. doi:10.2307/2344614

SAS Institute. 2012. SAS OnlineDoc 9.3. SAS Inst., Cary, NC.

Schabenberger, O., and F.J. Pierce. 2002. Contemporary statistical models for the plant and soil sciences. CRC Press, Boca Raton, FL.

Searle, S.R. 1971. Linear models. John Wiley & Sons, New York.

Snedecor, G., and W.G. Cochran. 1989. Statistical methods. 8th ed. Iowa State Univ. Press, Ames.

Stroup, W.W. 2013a. Generalized linear mixed models. CRC Press, Boca Raton, FL.

Stroup, W.W. 2013b. Non-normal data in agricultural experiments. In: Proceedings of the 25th Annual Conference on Applied Statistics in Agriculture, Manhattan, KS. 28–30 Apr. 2013. Kansas State Univ., Manhattan (in press).

University Statisticians of Southern Experiment Stations. 1989. Applications of mixed models in agriculture and related disciplines. South. Coop. Ser. Bull. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Wolfinger, R.D., and M. O'Connell. 1993. Generalized linear mixed models: A pseudo-likelihood approach. J. Stat. Comput. Simul. 48:233–243. doi:10.1080/00949659308811554

Yates, F. 1935. Complex experiments. J. R. Stat. Soc. 2(Suppl.):181–223.

Yates, F. 1940. The recovery of inter-block information in balanced incomplete block designs. Ann. Eugen. 10:317–325. doi:10.1111/j.1469-1809.1940.tb02257.x

Zeger, S.L., K.-Y. Liang, and P.S. Albert. 1988. Models for longitudinal data: A generalized estimating equation approach. Biometrics 44:1049–1060. doi:10.2307/2531734