Published 2012

CHAPTER 7

DESIGNING EXPERIMENTS

7.1 INTRODUCTION

In this chapter the use of generalized linear mixed models as a planning tool for the design of agronomic experiments is discussed. The reader might well ask, "Don't generalized linear mixed models concern modeling and data analysis? What does this have to do with design?" To quote from the classic text *Experimental Designs* (Cochran and Cox, 1992), "It has come to be recognized that the time to think about statistical inference, or to seek [a statistician's] advice, is when the experiment is being planned." Hahn (1984) put it more forcefully, "Statisticians make their most valuable contributions if they are consulted in the planning stages of an investigation. Proper experimental design is often more important than sophisticated statistical analysis." He continues, quoting H. Ginsburg as saying, "When I'm called in after it's all over, I often feel like a coroner. I can sign the death certificate—but do little more." Light et al. (1990) stated it slightly differently, "You cannot save by analysis what you bungle by design."

In his text *The Design of Experiments*, Mead (1988) noted that the development of experimental design concepts was "restricted by the earlier need to develop mathematical theory for design in such a way that the results from the design can be analyzed without recourse to computers." Because of the increasing sophistication of statistical modeling and the dramatic increase in capacity of modern computers, Mead argued, "The fundamental concepts now require reexamination and re-interpretation outside the limits of classical mathematical theory so that the full range of design possibilities may be considered."

Following his line of thought, while generalized linear mixed models provide researchers with expanded flexibility to apply regression and analysis of variance approaches to data that are not normally distributed, conventional wisdom about the design of experiments reflects the "restraints" referred to by Mead. For researchers to genuinely benefit from generalized linear mixed models, experiments must be designed to allow their full potential to be realized. This is done by using generalized linear mixed model power, precision, and sample-size analysis in the planning process.

doi:10.2134/2012.generalized-linear-mixed-models.c7

American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America
5585 Guilford Road, Madison, WI 53711-5801, USA.

Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences
Edward E. Gbur, Walter W. Stroup, Kevin S. McCarter, Susan Durham, Linda J. Young, Mary Christman, Mark West, and Matthew Kramer

As an example of an area where this type of pre-experiment preparation is rigorously followed, consider the pharmaceutical industry. Regulations require that investigators finalize study protocols before their commencement. A protocol must describe the design of the study, identify and rank, in order of importance, the various hypotheses to be tested, and specify the models to be fit and the statistical methods to be used in performing the analyses. As part of these preparations, power analyses are conducted to ensure that the study will be adequate for its intended purpose. This is very important. Even aside from financial considerations, it would be unethical to expose subjects to the potential risks of a clinical trial without ensuring a reasonable chance of detecting a clinically relevant treatment effect. In addition, it is undesirable to expose more subjects to the potential risks than are necessary to obtain a specified level of power.

This level of pre-experiment preparation is not, and may never be, required of researchers in most academic fields. However, it can be considered a "best practice model," a goal to strive for. In fact, we are seeing a movement in this direction in several fields. For example, grant-funding agencies such as NIH now require that power analyses be included in grant proposals. Even when not formally required, including a power analysis gives a grant proposal a competitive advantage because it shows funding agencies that the researcher has thought carefully about the proposed design and its potential to obtain results. In all cases, it is in the researcher's enlightened self-interest to assess the power and precision of a proposed design before data collection begins. This is especially true when generalized linear mixed models are to be used to analyze the data. A design that is optimal for analysis of variance or regression with normally distributed data may be unsuitable for non-normal data such as counts, percentages, and times to an event. What reasonable researcher would invest time, effort, and money in an experiment without first getting an idea of the likelihood of successfully detecting scientifically relevant results, should they exist?

The purpose of this chapter is to show how generalized linear mixed model based tools can be used in planning experiments that will be analyzed using generalized linear mixed models. Specifically, we show how generalized linear mixed models can be used to assess the expected power profile and the precision of a proposed experiment of a given size and type, and to guide modifications when they are necessary. In many cases, a given set of treatments and a given number of experimental units can be arranged into more than one plausible design, often with very different power profiles with respect to the researcher's objectives. Power and precision analysis can be used to assess the strengths and drawbacks of competing designs. The tools presented in this chapter should be considered essential in planning agronomic experiments and experiments in other fields as well.

7.2 POWER AND PRECISION

Power is defined as the probability of rejecting the null hypothesis when in fact the null hypothesis is false and therefore should be rejected. In practical terms, the null hypothesis states that a given treatment has no effect, while the research or alternative hypothesis states that a treatment does indeed have an effect. Hence, power is the probability that one will be able to demonstrate the credibility of the research hypothesis, with acceptable scientific rigor, when the research hypothesis is in fact true.

Power analysis is, in essence, the computation of that probability. Specifically, one determines the minimum treatment effect one considers to be scientifically relevant and then computes the probability that a proposed design will show that difference to be statistically significant. Precision analysis is similar, but instead of focusing on power, one determines how wide a confidence interval for the treatment effect is expected to be for the proposed design.

7.3 POWER AND PRECISION ANALYSES FOR GENERALIZED LINEAR MIXED MODELS

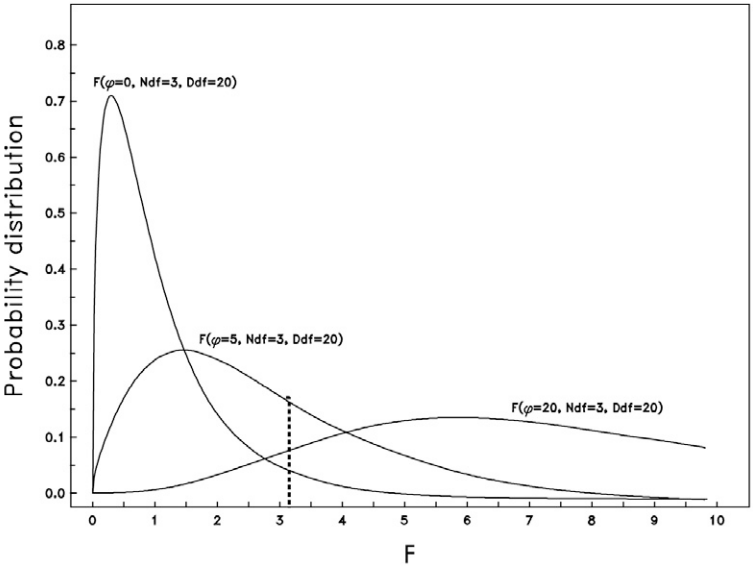
The first step in conducting a power and a precision analysis is to identify the nature of the response variable, its distribution, and the effect(s) of interest. For example, in a one factor, completely randomized design, the model describing the treatment effect is $\beta_0 + T_i$, $i = 1, \ldots, t$, where T_i is the effect of the ith treatment and β_0 is the intercept or overall mean. For normally distributed response variables, $\beta_0 + T_i$ directly models the treatment mean μ_i. For binomial responses, $\beta_0 + T_i$ usually models the logit of π_i, where π_i denotes the probability of the occurrence of the event of interest (success) for the ith treatment. For counts modeled by an appropriate counting distribution, $\beta_0 + T_i$ models $\log(\mu_i)$, where μ_i is the expected count for the ith treatment.

The hypotheses to be tested are specified in terms of treatment differences or, more generally, contrasts $\Sigma_i k_i T_i$, where the k_i are constants chosen to define the effect of interest. Under the null hypothesis, H_0: $\Sigma_i k_i T_i = 0$ and under the research hypothesis, H_A: $\Sigma_i k_i T_i \neq 0$. For example, setting $k_1 = 1$, $k_2 = -1$, and the remaining $k_i = 0$ defines the contrast $T_1 - T_2$, the difference between treatments 1 and 2. In this case, H_0: $T_1 - T_2 = 0$ and H_A: $T_1 - T_2 \neq 0$. A generalized linear mixed model test of this hypothesis is based on an F statistic. If H_0 is true, this statistic has an approximate central F distribution, denoted $F_{(0,\ \text{Ndf},\ \text{Ddf})}$, where Ndf denotes the numerator degrees of freedom, and Ddf denotes the denominator degrees of freedom. Under the research hypothesis, the F statistic has an approximate non-central F distribution, denoted by $F_{(\varphi,\ \text{Ndf},\ \text{Ddf})}$, where φ denotes the non-centrality parameter. Without going into technical details, the non-centrality parameter depends on the quantity

$$\left(\text{sample size}\right) \times \left(\sum_i k_i T_i\right) \Big/ \left(\text{variance of treatment effect}\right)$$

A formal definition and technical details can be found in experimental design textbooks, for example, Hinkelmann and Kempthorne (1994). Note that under the null hypothesis, $\Sigma_i k_i T_i = 0$, and hence, the non-centrality parameter φ is also 0. Under the research hypothesis, $\Sigma_i k_i T_i > 0$, and hence, $\varphi > 0$. The non-centrality parameter

FIG. 7–1. The effect of the non-centrality parameter on the F distribution.



increases when either the effective sample size increases, the treatment effect increases, or the variance of the treatment effect decreases.

Figure 7–1 illustrates the effect of the non-centrality parameter on the F distribution. In the figure, the central F is the highly right-skewed distribution in the left-most position and represents the distribution of the test statistic under the null hypothesis. The dashed vertical line represents the critical value of the test for $\alpha = 0.05$. An observed value of the F statistic greater than the critical value would lead to rejection of the null hypothesis. The two non-central F distributions show what happens as the non-centrality parameter increases; namely, the larger the value, the more the distribution is shifted to the right. The area under the curve to the right of the critical value corresponds to the power of the test. As φ increases, the power of the test increases.

Precision analysis is based on interval estimation of the effect of interest, $\Sigma_i k_i T_i$. The ratio of the estimated contrast to its standard error has an approximate t distribution with Ddf degrees of freedom. Thus, a $100(1 - \alpha)\%$ confidence interval for $\Sigma_i k_i T_i$ is of the form

$$\text{estimate of} \sum_i k_i T_i \pm t_{(\alpha,\ \mathrm{Ddf})} \times \text{standard error of} \sum_i k_i T_i$$

For a given design, one can use generalized linear mixed model software to compute the approximate standard error and hence, the expected confidence interval width for the contrast.

7.4 METHODS OF DETERMINING POWER AND PRECISION

There are two primary ways of evaluating the power and precision of an experiment using generalized linear mixed model software. The first method, henceforth referred to as the probability distribution method, is applicable when we know (or can approximate) the sampling distribution of the test statistic under the conditions of the research hypothesis. In this case, one determines the non-centrality parameter of the distribution of the test statistic at a particular point under the research hypothesis. One then approximates the power of the test using the area under this non-central distribution to the right of the critical value, as illustrated in Fig. 7–1. One can use GLIMMIX in conjunction with SAS's (SAS Institute, Cary, NC) probability functions to perform these calculations.

The second method uses simulation to estimate the power of a test and is applicable regardless of whether we know or can approximate the actual sampling distribution of the test statistic under the research hypothesis. All that is necessary to use this method is the ability to perform the test of interest and generate random numbers from the distribution of interest. To estimate power via simulation, one uses a random number generator to create a large number of independent data sets that match the proposed study design and reflect the conditions under the research hypothesis to be detected. The SAS data step and random number functions can be used to create these data sets. One then performs the desired generalized linear mixed model analysis for each data set separately, in each case keeping track of whether the null hypothesis has been rejected. Since the simulated samples are independent, the number of samples for which the null hypothesis is rejected has a binomial distribution with the number of trials equal to the number of simulated datasets and with probability equal to the true power of the hypothesis test. This fact provides a basis for making inferences about the true power of the test, including computing point and confidence interval estimates and testing hypotheses about the power. In particular, the proportion of simulated datasets for which the null hypothesis is rejected gives a point estimate of the power of the test. In addition, a confidence interval for the power of the test can be computed from these simulation results. For precision analysis, one calculates the mean and variance of the width of the confidence intervals over all simulated samples produced by the generalized linear mixed model. GLIMMIX can be used for both power analysis and precision analysis using the simulation method.

A major advantage of the probability distribution method over the simulation method is that it is quicker and easier to set up, allowing rapid comparison of competing designs, different effect sizes, different levels of variation, or different sample sizes. However, to use this method one must know (or know an approximation of) the actual sampling distribution of the test statistic. One advantage of the simulation method over the probability distribution method is that it is applicable for any design and any type of analysis, regardless of whether the behavior of the test statistic is well understood. The only requirement is that one is able to generate data according to the study design.

A second advantage of the simulation method is that, since it involves analyzing hundreds (or thousands) of datasets similar to those that are expected from the study, it allows one to see exactly what the analysis will look like and how the GLIMMIX procedure will behave with data from the proposed design. The simulation method may reveal any troublesome behavior GLIMMIX may display for a contemplated design. Researchers can use such fair warning to make needed changes in the proposed design before the data are collected and it is too late. One disadvantage of the simulation method is that, because it requires analyzing a large number of samples, it can be much more time consuming, especially when evaluating power over a wide range of possibilities under the alternative hypothesis. A benefit of both methods is that the programs used to perform the power analysis can be used later, perhaps with minor alterations, to analyze the real data once they have been obtained.

The approach taken in this chapter and recommended for use in practice is to use the probability distribution method to compare the various design alternatives for a study and to identify one or more that provide the desired power characteristics. Then use the simulation method to verify the power approximations obtained from the probability distribution method. Again we emphasize that all of this should be done during the planning stages of an experiment, before data collection starts.

Four items of information are required to perform a power analysis:

- the minimum treatment effect size $\Sigma_i k_i T_i$ considered scientifically relevant,
- the assumed probability distribution of the response variable,
- an approximate idea of the magnitude and nature of the variation and correlation present in the data,
- a clear idea of the structure of the proposed design.

A few clarifications about these required items are in order. First, providing the scientifically relevant treatment effect size does not mean knowing in advance how big a difference there will be among treatment means. Many researchers short-circuit power analysis by saying, "I can't give you that. If I knew how different the treatment means are, I wouldn't have to run the experiment!" True, but that is not the question. The question is, "Given your knowledge about the research question that is motivating this study, what is the minimum difference that would be considered important if, in fact, it exists?" Would a 1 kg ha^{-1} increase in yield be considered too trivial to matter? Would a 10 kg ha^{-1} difference be considered extremely important? What about a 5 kg ha^{-1} difference?

Second, to get an idea of the magnitude and nature of the variation and correlation present in the data, one must identify the relevant sources of variation (e.g., blocking, experimental unit error), distinguishing between whole plot and split plot variance if a split plot experiment is being proposed, characterizing likely correlation structures among measurements over time if a repeated measures design is proposed, and characterizing likely spatial variability if there is reason to believe it is present. Several of these issues will be addressed in the examples that follow.

If it appears from these requirements that a great deal of conversation between the researcher and statistical scientist should be occurring early in the planning of the experiment, then the reader has the right idea.

7.5 IMPLEMENTATION OF THE PROBABILITY DISTRIBUTION METHOD

This basic approach originated with linear models using PROC GLM in SAS (Littell, 1980; O'Brien and Lohr, 1984; Lohr and O'Brien, 1984). Stroup (1999) extended the method to linear mixed models using PROC MIXED. Stroup (2002) described the implementation of the probability distribution method for linear mixed models using PROC MIXED, focusing on experiments in the presence of spatial variation, and provided evidence of the accuracy of these methods via simulation. Littell et al. (2006) provided additional detail and examples for linear mixed models. In this section, the method is extended to generalized linear mixed models.

Implementation of the probability distribution method requires four basic steps. These steps are listed here and are illustrated by a simple example using a two-treatment, completely randomized design for a normally distributed response. The steps are as follows:

1. Create an "exemplary data set" (O'Brien and Lohr, 1984), that is, a data set whose structure is identical to the data that would be collected using the proposed design but with the observed data replaced by means reflecting the treatment difference to be detected under the research hypothesis.
2. Determine the numerator and denominator degrees of freedom and the non-centrality parameter that follows from the design and the research hypothesis. These can be obtained from the generalized linear mixed model software.
3. Determine the critical value based on the numerator and denominator degrees of freedom found in Step 2.
4. Compute the power, that is, the probability that the test statistic exceeds the critical value using the numerator and denominator degrees of freedom and the non-centrality parameter determined in Step 2 and the critical value found in Step 3.

EXAMPLE 7.1

Suppose we want to compare two treatments, a reference (or control) and an experimental treatment, using a completely randomized design in which the response is normally distributed. Suppose further that experience with the control treatment indicates it has a mean response of approximately 10 units with a standard deviation of roughly 10% of the mean. That is, for the control treatment $\mu = 10$ and $\sigma = 1$. The researcher believes that it would be scientifically relevant if the experimental treatment increases the mean response by 10% or more; i.e., to at least $\mu = 11$. The researcher wants to know the probability that four replications per treatment would show the scientifically relevant difference to be statistically significant. With this information, the probability distribution method is implemented as follows.

FIG. 7–2. SAS statements to create an exemplary data set for Example 7.1.

```
data crd_example;
   input trt mu;
   do rep=1 to 4 by 1;
     output;
   end;
   datalines;
 0  10
 1  11
run;

proc print data=crd_example noobs;
   var trt mu rep;
run;
```

FIG. 7–3. The exemplary data set for Example 7.1 from the PROC PRINT in Fig. 7–2.

```
 trt      mu      rep
 0        10       1
 0        10       2
 0        10       3
 0        10       4
 1        11       1
 1        11       2
 1        11       3
 1        11       4
```

Step 1. Create the exemplary data set.

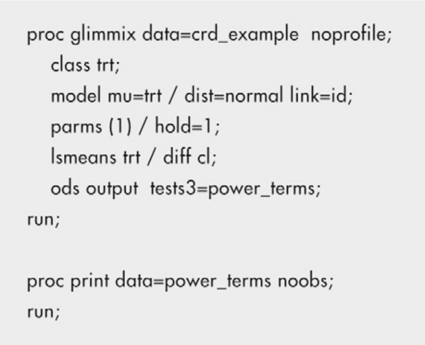
This will have four lines of data per treatment (one per replication), each line containing the treatment and the mean for that treatment under the research hypothesis (10 for control, 11 for the experimental treatment).

The SAS data step to accomplish Step 1 is shown in Fig. 7–2, and the data file that it created is shown in Fig. 7–3. There are two input variables, *trt* (treatment) and *mu* (the mean for the treatment specified by *trt*). *trt* takes two values, 0 for the control and 1 for the experimental treatment, and *mu* takes the values 10 and 11, respectively, corresponding to the minimum scientifically relevant difference as specified by the researcher. The *do*, *output*, and *end* statements form a "do-loop" to create the required four lines of data per treatment.

Step 2. Analyze the exemplary dataset using GLIMMIX to obtain the terms needed to compute the power and the precision of the experiment.

The GLIMMIX statements for Step 2 are given in Fig. 7–4. The *class* and *model* statements are exactly as they would be when the actual data from the experiment are analyzed. The *parms* statement sets the error variance to 1. The *hold* option instructs the procedure to fix it at $\sigma^2 = 1$ and to not treat it as a parameter to be estimated. (The *parms* statement and *noprofile* option would be removed when analyzing the real data). The *diff* and *cl* options in the LSMEANS statement direct the procedure to compute the projected 95% confidence interval for the treatment difference. For this example, this is the precision analysis. The output shown in Fig. 7–5 gives the information needed for the precision analysis. The *ods* statement causes the GLIMMIX procedure to create a new data set, which we have named *power_terms*, that contains the various values needed for the power analysis (F value, numerator and denominator degrees of freedom). The contents of this file are shown in Fig. 7–6.

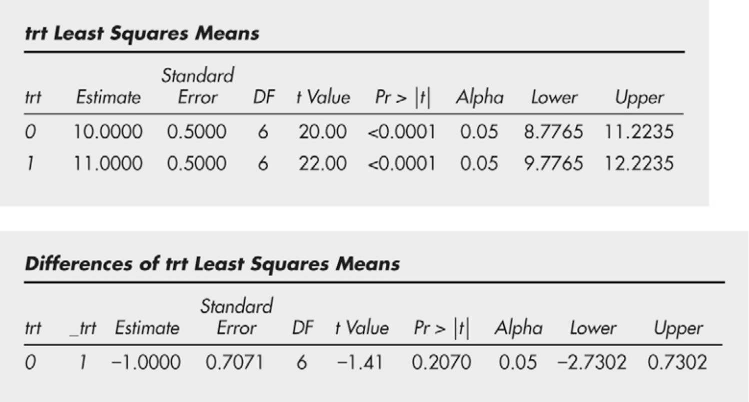
FIG. 7–4. GLIMMIX statements to compute terms needed for the power/precision analysis for Example 7.1.



```
proc glimmix data=crd_example  noprofile;
   class trt;
   model mu=trt / dist=normal link=id;
   parms (1) / hold=1;
   lsmeans trt / diff cl;
   ods output  tests3=power_terms;
run;

proc print data=power_terms noobs;
run;
```

FIG. 7–5. GLIMMIX output containing the information required for the precision analysis in Example 7.1.



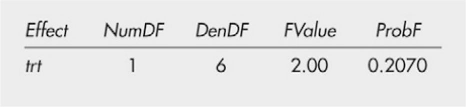
```
trt Least Squares Means

              Standard
trt  Estimate   Error   DF  t Value  Pr > |t|  Alpha   Lower    Upper
0    10.0000   0.5000   6   20.00    <0.0001   0.05   8.7765   11.2235
1    11.0000   0.5000   6   22.00    <0.0001   0.05   9.7765   12.2235
```

```
Differences of trt Least Squares Means

                       Standard
trt  _trt  Estimate     Error   DF  t Value  Pr > |t|  Alpha    Lower    Upper
0       1  -1.0000     0.7071   6   -1.41    0.2070    0.05   -2.7302   0.7302
```

FIG. 7–6. The contents of the *power_terms* file for Example 7.1 from the PROC PRINT in Fig. 7–4.



```
Effect   NumDF   DenDF   FValue   ProbF
trt          1       6     2.00   0.2070
```

From Fig. 7–5, the precision analysis shows that if this experiment is run with four replications, the expected standard error of a treatment mean will be 0.5, the expected standard error of a treatment difference will be 0.707, and the expected width of the 95% confidence interval for the treatment difference will be 0.730 − (−2.730) = 3.46 units.

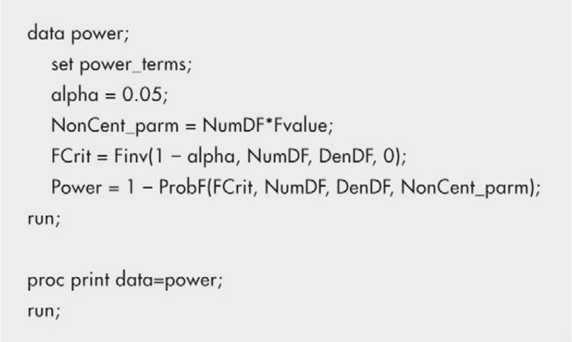
Steps 3 and 4. The values in the data set created by the *ods* statement (*power_terms*) are used to obtain the critical value, compute the non-centrality parameter, and then evaluate the power.

The SAS statements to perform Steps 3 and 4 are shown in Fig. 7–7. These statements, perhaps with minor alterations, are used for all of the examples presented in this chapter. The data step creates a new data set called *power* from the data set *power_terms* produced by the GLIMMIX analysis. The non-centrality parameter under the research hypothesis is equal to the product of the numerator degrees of freedom (*NumDF*) and the F-value. In this example, α, the type I error probability, is set to 0.05. The critical value of F is calculated using the *finv* function. The statement shown obtains the critical value from the central F-distribution (i.e., F under the null hypothesis) using the numerator and denominator degrees of freedom provided by GLIMMIX as a result of analyzing the exemplary dataset. The *ProbF* function determines the area under the non-central F distribution (i.e., F under the research hypothesis) to the left of the critical value. Subtracting this area from one yields the power. The resulting information from the PROC PRINT statement appears in Fig. 7–8.

The approximated power of the proposed experiment is 0.2232. In other words, given the scientifically relevant difference specified above and the assumed magnitude of the error variance, the researcher has less than a one in four chance of obtaining data that will allow rejection of the null hypothesis. Clearly, four replications do not provide adequate power.

One can evaluate power for different numbers of replications by modifying the upper limit in the *do* statement in the creation of the exemplary data set. To find the minimum number of replications required to obtain a given power, one can

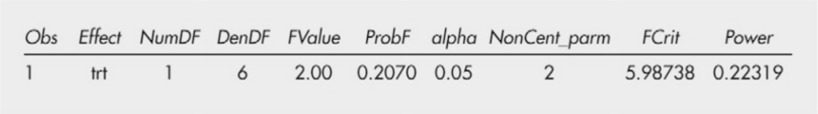
FIG. 7–7. SAS statements to compute power from the GLIMMIX output for Example 7.1.



```
data power;
   set power_terms;
   alpha = 0.05;
   NonCent_parm = NumDF*Fvalue;
   FCrit = Finv(1 – alpha, NumDF, DenDF, 0);
   Power = 1 – ProbF(FCrit, NumDF, DenDF, NonCent_parm);
run;

proc print data=power;
run;
```

FIG. 7–8. Power analysis results for Example 7.1 from the PROC PRINT in Fig. 7–5.



```
Obs  Effect  NumDF  DenDF  FValue  ProbF   alpha  NonCent_parm  FCrit    Power
1    trt     1      6      2.00    0.2070  0.05   2             5.98738  0.22319
```

progressively change this upper limit until the desired level of power is obtained. For example, suppose we wish to determine the smallest number of replications for which the test has power at least 0.80. Varying the upper endpoint in the *do* statement in this way, we find that 16 replications result in the power being 0.78, and 17 replications result in the power being 0.81. Therefore 17 is the minimum number of replications that will provide at least an 80% chance of detecting the treatment difference specified above, assuming an error variance of 1. The power provided by other numbers of replications when the error variance is 1 is given in Table 7–1 under the column labeled approximated power.

TABLE 7–1. Approximated and estimated power for the comparison of two treatments in a completely randomized design with variance equal to 1 in Example 7.1.

Number of replications	Approximated power†	Estimated power†	Lower confidence limit‡	Upper confidence limit‡
4	0.2232	0.2188	0.1934	0.2441
10	0.5620	0.5713	0.5410	0.6016
15	0.7529	0.7813	0.7559	0.8066
16	0.7814	0.7813	0.7559	0.8066
17	0.8070	0.7881	0.7631	0.8131
18	0.8300	0.8379	0.8153	0.8605
19	0.8506	0.8721	0.8516	0.8925
20	0.8690	0.8652	0.8443	0.8861
25	0.9337	0.9287	0.9130	0.9445
30	0.9677	0.9648	0.9536	0.9761
31	0.9721	0.9678	0.9570	0.9786
32	0.9760	0.9746	0.9650	0.9842
33	0.9793	0.9678	0.9570	0.9786
34	0.9822	0.9795	0.9708	0.9882
35	0.9848	0.9834	0.9756	0.9912
40	0.9930	0.9941	0.9895	0.9988
45	0.9968	0.9990	0.9971	1.0000
50	0.9986	0.9990	0.9971	1.0000
55	0.9994	0.9961	0.9923	0.9999
60	0.9997	0.9990	0.9971	1.0000
65	0.9999	1.0000	1.0000	1.0000
70	1.0000	1.0000	1.0000	1.0000

† Approximated power is based on the probability distribution method. Estimated power is based on the simulation method with 1024 simulated samples.

‡ 95% confidence limits for the estimated power.

What is the effect on power if we have underestimated the error variance? For example, how much power will 17 replications provide if the error variance σ^2 is actually 2, or even worse if it is as large as 4? This is easily answered by changing the *parms* statement in the GLIMMIX procedure. Re-running the procedure above with *parms (2)*, we see that with 17 replications the power drops to 0.52 when $\sigma^2 = 2$ (Table 7–2), and re-running with *parms (4)* shows that with 17 replications the power drops even further to 0.29 when $\sigma^2 = 4$ (Table 7–3). By increasing the number of replications as described above we see that if σ^2 were actually 4, it would take 65 replications to achieve power of 0.80 (Table 7–3).

TABLE 7–2 Approximated and estimated power for the comparison of two treatments in a completely randomized design with variance equal to 2 in Example 7.1.

Number of replications	Approximated power†	Estimated power†	Lower confidence limit‡	Upper confidence limit‡
4	0.1356	0.1357	0.1148	0.1567
10	0.3220	0.3408	0.3118	0.3699
15	0.4642	0.4551	0.4246	0.4856
16	0.4904	0.4932	0.4625	0.5238
17	0.5158	0.5029	0.4723	0.5336
18	0.5403	0.5547	0.5242	0.5851
19	0.5640	0.5684	0.5380	0.5987
20	0.5868	0.5908	0.5607	0.6209
25	0.6879	0.6885	0.6601	0.7168
30	0.7682	0.7578	0.7316	0.7841
31	0.7820	0.7715	0.7458	0.7972
32	0.7951	0.8008	0.7763	0.8252
33	0.8076	0.8008	0.7763	0.8252
34	0.8193	0.8135	0.7896	0.8373
35	0.8305	0.8262	0.8030	0.8494
40	0.8776	0.8711	0.8506	0.8916
45	0.9127	0.9199	0.9033	0.9365
50	0.9383	0.9424	0.9281	0.9567
55	0.9568	0.9531	0.9402	0.9661
60	0.9700	0.9678	0.9570	0.9786
65	0.9794	0.9795	0.9708	0.9882
70	0.9859	0.9873	0.9804	0.9942

† Approximated power is based on the probability distribution method. Estimated power is based on the simulation method with 1024 simulated samples.

‡ 95% confidence limits for the estimated power.

TABLE 7–3 Approximated and estimated power for the comparison of two treatments in a completely randomized design with variance equal to 4 in Example 7.1.

Number of replications	Approximated power†	Estimated power†	Lower confidence limit‡	Upper confidence limit‡
4	0.0923	0.0908	0.0732	0.1084
10	0.1851	0.1943	0.1701	0.2186
15	0.2624	0.2695	0.2424	0.2967
16	0.2777	0.2803	0.2528	0.3078
17	0.2930	0.2871	0.2594	0.3148
18	0.3081	0.3018	0.2736	0.3299
19	0.3231	0.3281	0.2994	0.3569
20	0.3379	0.3398	0.3108	0.3689
25	0.4101	0.4102	0.3800	0.4403
30	0.4779	0.4873	0.4567	0.5179
31	0.4909	0.4785	0.4479	0.5091
32	0.5036	0.4922	0.4616	0.5228
33	0.5162	0.5449	0.5144	0.5754
34	0.5285	0.5146	0.4840	0.5453
35	0.5407	0.5352	0.5046	0.5657
40	0.5981	0.6074	0.5775	0.6373
45	0.6502	0.6523	0.6232	0.6815
50	0.6969	0.7158	0.6882	0.7434
55	0.7385	0.7393	0.7124	0.7661
60	0.7753	0.7773	0.7519	0.8028
65	0.8076	0.8057	0.7814	0.8299
70	0.8358	0.8525	0.8308	0.8743

† Approximated power is based on the probability distribution method. Estimated power is based on the simulation method with 1024 simulated samples.

‡ 95% confidence limits for the estimated power.

One could also change the variance in the PARMS statement to find the largest variance for which four replications provide an 80% chance of detecting the treatment difference. Doing this, we find that the largest the variance can be in this case is $\sigma^2 = 0.18$. Finally, one could modify the variable *mu* to determine the minimum treatment difference four replications could detect at a significance level of $\alpha = 0.05$ and power = 0.80, with $\sigma^2 = 1$. For example, *mu* = 10 and 12.4 for *trt* = 0 and 1, respectively (a 24% difference), yields a power of 0.806.

The power calculations above are based on a generalized linear mixed model analysis, and therefore are based on an F statistic. The F distribution is an approximation to the true sampling distribution of the generalized linear mixed model test statistic. Hence, the power values obtained above are approximations as well. We can assess the accuracy of these approximations by estimating the true power using the simulation method discussed in Section 7.4. To do this, 1024 independent random samples were generated according to the model using the same combinations of assumed variance and number of blocks considered in the probability distribution method above. Each sample was analyzed using the GLIMMIX model shown in Fig. 7–4 (excluding the *parms* statement and the *noprofile* option). The results of the analyses of the simulated samples were used to calculate point and confidence interval estimates of the true power for each number of blocks under consideration. As can be seen in Tables 7–1, 7–2, and 7–3, in most cases the approximated power values are contained within the 95% confidence interval estimates of the true power. From this we can conclude that the power approximations obtained from the probability distribution method are accurate in this scenario. This illustrates the general result that for response variables with a normal distribution, this approximation is very good, and therefore we can be confident in the results provided by the probability distribution method in such cases. For response variables with non-normal distributions, using simulation to verify the results obtained from the probability distribution method is much more important because the non-normal case has not been studied as extensively and less is known about its performance in certain cases, such as when the number of replications is small. ■

This simple example demonstrates the use of the probability distribution and simulation methods for evaluating power and precision. The remaining examples show how these methods can be used to perform power and precision analysis for several more realistic situations involving generalized linear mixed models.

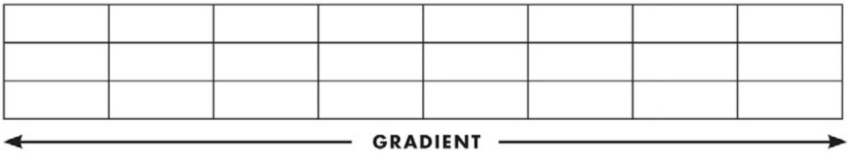
7.6 A FACTORIAL EXPERIMENT WITH DIFFERENT DESIGN OPTIONS

The example in this section shows three alternative ways of setting up a two-factor factorial experiment with a given set of experimental units. Each design exhibits different power and precision characteristics, thereby providing the scientist with choices on ways to obtain more information from a fixed set of resources.

EXAMPLE 7.2

A researcher wants to conduct a field experiment to compare two treatments at three rates of application. For example, the two treatments could be two methods of application, two tillage methods, or two varieties. The three rates of application could represent amounts of a fertilizer or pesticide or irrigation levels. Treatment designs identical or similar to this two-treatment × three-rate factorial occur frequently in agronomic research. Assume that the response is normally distributed.

FIG. 7–9. Field layout of the experimental plots for a 3 × 2 factorial treatment structure for Examples 7.2, 7.5, 7.6, and 7.7.



Now suppose the resources available to the researcher consist of an 8 × 3 grid of plots with a gradient parallel to the direction of the 8-plot rows. Figure 7–9 shows the field layout.

The variation among the three-plot columns due to the gradient suggests that some form of blocking is advisable. Since there are six treatment × rate combinations in the treatment design, one obvious blocking strategy would combine pairs of adjacent columns into blocks, resulting in a randomized complete block (RCB) design with four blocks. However, with a strong enough gradient, adjacent columns may be dissimilar, resulting in excessively heterogeneous experimental units within blocks, a well-known poor design idea. An alternative design would use each 3-plot column as an incomplete block and set the experiment up as an incomplete block (IB) design with 8 blocks. A third approach would be to form blocks as in the randomized complete block design, assigning treatments to 3-plot columns within a block (whole plots), and then randomly assigning rates to subplots within each whole plot, resulting in a split plot (SP) design with an RCB whole plot design structure. Figure 7–10 shows a layout for each design.

Each design in Fig. 7–10 requires a different model for analysis, resulting in potentially different power characteristics. Each model consists of a component related to the treatment structure and a component related to the design structure of the experiment (Milliken and Johnson, 2009). The treatment structure is the same for all three experiments. Each model has in common the treatment × rate structure given by

$$\mu_{ij} = \beta_0 + T_i + D_j + \mathrm{TD}_{ij}$$

where μ_{ij} is the mean for the ith treatment and jth rate, β_0 is the intercept, T_i is the ith treatment effect, D_j is the jth rate effect, and TD_{ij} is the effect of the treatment × rate interaction. The models differ in their blocking and error structures, which make up the remainder of each model. The complete models are given below.

- Randomized complete block (RCB):

  $$Y_{ijk} = \mu_{ij} + R_k + e^{\mathrm{R}}_{ijk}$$

  where R_k is the kth block effect, assumed to be independent $N(0, \sigma^2_{\mathrm{R}})$, and e^{R}_{ijk} is the error term, assumed to be independent $N(0, \sigma^2_{e\mathrm{R}})$. The subscript and superscript R denotes the randomized complete block design.

FIG. 7–10. Field layouts as a randomized complete block, an incomplete block, and a split plot with whole plots in blocks for Example 7.2.

Randomized complete block

Block 1		*Block 2*		*Block 3*		*Block 4*	
T2 D1	T2 D3	T1 D3	T1 D2	T2 D1	T1 D1	T2 D1	T1 D2
T1 D3	T2 D2	T1 D1	T2 D3	T2 D2	T2 D3	T1 D3	T2 D3
T1 D2	T1 D1	T2 D1	T2 D2	T1 D3	T1 D2	T2 D2	T1 D1

Incomplete block

Block 1	*Block 2*	*Block 3*	*Block 4*	*Block 5*	*Block 6*	*Block 7*	*Block 8*
T2 D3	T2 D1	T2 D3	T2 D2	T2 D1	T2 D2	T2 D3	T2 D3
T1 D2	T1 D3	T1 D3	T2 D1	T1 D3	T1 D3	T2 D1	T2 D2
T1 D1	T1 D1	T1 D1	T1 D1	T1 D2	T1 D2	T1 D2	T1 D3

Split plot with whole plot in blocks

Block 1		*Block 2*		*Block 3*		*Block 4*	
T1 D1	T2 D2	T2 D2	T1 D3	T2 D3	T1 D1	T1 D2	T2 D3
T1 D3	T2 D1	T2 D3	T1 D2	T2 D1	T1 D2	T1 D1	T2 D2
T1 D2	T2 D3	T2 D1	T1 D1	T2 D2	T1 D3	T1 D3	T2 D1

- Incomplete block design (IB):

  $$Y_{ijk} = \mu_{ij} + B_k + e^{\mathrm{I}}_{ijk}$$

  where B_k is the kth incomplete block effect, assumed to be independent $N(0, \sigma^2_B)$, and e^{I}_{ijk} is the error term, assumed to be independent $N(0, \sigma^2_{e\mathrm{I}})$. The subscript and superscript I denotes the incomplete block design.

- Split plot (SP):

  $$Y_{ijk} = \mu_{ij} + R_k + w_{ik} + s_{ijk}$$

  where R_k is the kth block effect (as in the RCB), w_{ik} is the whole plot error, assumed to be independent $N(0, \sigma^2_W)$, and s_{ijk} is the split plot error, assumed to be independent $N(0, \sigma^2_S)$.

Once we specify the plausible designs and their associated models, we have a decision to make. Which one should the researcher use? Assuming that the design costs are the same for the above designs, the answer is the design that maximizes power and precision for the treatment comparisons that address the researcher's objectives. To do the power analysis required to make this determination, we need to specify values of the μ_{ij} under the research hypothesis; that is, what are the "agronomically relevant" differences among these treatments and what comparisons among them best address the objectives?

FIG. 7–11. SAS statements to create an exemplary data set for the split plot design for Example 7.2.

```
data exemplary_data;
   input trt rate mu;
   do block=1 to 4 by 1;
      output;
   end;
datalines;
 1  1  105.72
 1  2  111.44
 1  3  117.16
 2  1  111.44
 2  2  122.88
 2  3  134.32
run;
```

As an example, suppose treatment 1 historically showed a 5.72 unit increase each time the rate was increased; e.g., from "low" to "medium" or from "medium" to "high." Suppose that the research hypothesis states that under treatment 2, the response to these rate increases would be greater. The researcher considered that a doubling of that rate would be "agronomically relevant." A power analysis for this research hypothesis is accomplished by performing a test of the equality of the linear effect of rate across each treatment, i.e., by testing the treatment × linear rate interaction. This hypothesis is tested in GLIMMIX using the following *contrast* statement:

*contrast 'trt × lin_rate' trt*rate −1 0 1 1 0 −1;*

Figure 7–11 shows the SAS data step used to create an exemplary data set for this power and precision analysis based on four blocks. The values for *mu* follow from the discussion above and the assumption that the mean response in the absence of any treatment is 100 units (any value could be used for this baseline).

Once the exemplary data set is specified, we need to specify the variance components associated with each design so that we can determine the non-centrality parameter. The variance structure in each model is a combination of the variance among plots within each column and the magnitude of the gradient. Suppose that enough is known about this structure to give the following information about the probable variance components that would result from each design.

- Randomized complete block (RCB): $\sigma_R^2 = 15$ and $\sigma_{e\mathrm{R}}^2 = 34$
- Incomplete block design (IB): $\sigma_B^2 = 35$ and $\sigma_{e\mathrm{I}}^2 = 14$
- Split plot (SP): $\sigma_R^2 = 15$, $\sigma_W^2 = 20$ and $\sigma_S^2 = 14$

Notice the difference between the RCB and IB variance components. The 3-plot columns are natural blocks induced by the gradient. The complete blocks are artificial "convenience" blocks constructed by combining natural blocks. Creating artificial blocks in this way reduces the variance among blocks and increases the error variance within blocks. This will affect power and precision.

Table 7–4 shows results from the precision analysis for these designs, specifically, the standard errors for various differences under each design. Note that the incomplete block design is best suited for comparisons between treatments (both main effects and simple effects at given rates), whereas the split plot design is best suited for comparisons among rates (split plot factor) but is least suited for comparisons among treatments (whole plot factor). For every effect, the randomized complete block design is less precise than the incomplete block.

TABLE 7–4. Precision analysis of competing designs for 3 × 2 factorial experiment in Example 7.2. Standard errors in bold indicate the best design for the corresponding effect.

	Design		
Effect	Randomized complete block	Incomplete block	Split plot
Treatment main effect	2.04	**1.80**	3.51
Rate main effect	2.50	1.98	**1.87**
Simple effect: treatment at *j*th rate	3.54	**2.99**	4.12
Simple effect: rate at *i*th treatment	3.54	2.86	**2.65**

FIG. 7–12. GLIMMIX statements to compute terms needed for the power/precision analysis for the split plot for Example 7.2.

```
proc glimmix data=exemplary_data  noprofile;
   class block trt rate;
   model mu=trt rate trt*rate / dist=normal  link=id;
   random intercept trt / subject=block;
   parms (15)(20)(14) / hold=1,2,3;
   contrast  'trt x lin_rate'  trt*rate  −1  0  1  1  0  −1;
   lsmeans  trt | rate / diff  slicediff=(trt  rate);
   ods output  contrasts=research_H_test_terms;
run;
```

Figure 7–12 shows the GLIMMIX statements needed to obtain the values for a power analysis for the split plot version of the experiment. The *random* statement accounts for the design structure of the experiment by incorporating random effects for the block and whole plot error terms. The *contrast* statement tests the research hypothesis of interest and produces values needed for the power analysis. The *ods* statement saves these values in a dataset we have called *research_h_test_terms*. These values are then processed as shown in Fig. 7–7 of Example 7.1. More than one *contrast* statement can be included and the *ods* statement can create multiple output data sets; for example, if one also wanted to output the type 3 test of fixed effects (*tests3*) results. The *lsmeans* statement produces results for the precision analysis.

By changing the *random* statements different design structures can be accommodated. This was done to calculate the power for the RCB and IB versions of the experiment as well. The power approximations for the three designs obtained using the probability distribution method are given below:

- Split plot (SP): approximated power 0.801
- Incomplete block (IB): approximated power 0.726

- Randomized complete block (RCB): approximated power 0.451

The three designs provide different levels of power. These results underline the take-home message of this example; namely, over-simplification of power analysis and sample size analysis often encourages misplaced focus in designing an experiment. To emphasize this point, consider the following scenario. Imagine that the researcher has done all the planning up to the point where the power is actually computed. The researcher, having had only a semester of statistical methods, is familiar only with randomized complete block designs and, therefore, has considered only that design. Shortly before the grant proposal is to be submitted, the researcher brings the statistician the information about the variance components and the agronomically relevant difference and asks for a power calculation, using the standard greeting, "I know you're busy, but I need this by noon today."

Once the power is computed, the statistician delivers the bad news. The power for four blocks is only 0.45. "How many blocks do I need to get the power up to 0.80?" By running the power algorithm above with different numbers of blocks, the statistician finds that nine blocks would be required. The researcher adjusts the budget to accommodate nine blocks and everyone lives happily ever after—except those whose money and labor have been wasted. The researcher has asked the wrong question. Rather than "How many blocks do I need?" the question should have been "What is the most efficient way to use the resources that I have available?" And, the researcher should also have asked this question much sooner. This conversation should have begun when the researcher was first thinking about this project. This scenario illustrates a point that should have particular resonance in a time of budget deficits, unpredictable energy costs, and tight money. ■

7.7 A MULTI-LOCATION EXPERIMENT WITH A BINOMIAL RESPONSE VARIABLE

This section illustrates another common experimental setting. From the statistical perspective, multi-location studies present the same basic statistical issues as laboratory studies conducted in multiple growth chambers or using other types of "identical" equipment or in studies conducted in multiple independent runs over time. In addition, some of the issues involved in designing experiments where the response of interest is a proportion are discussed. The considerations in these examples are applicable to any binomial response variable—dead/alive, damaged/undamaged, germinate/did not germinate, etc. There are standard textbook formulas for determining sample size with binomial response variables. However, as the examples will show, the standard formulas are inappropriate and inapplicable to the vast majority of agronomic experiments in which conclusions are to be based on binomial response variables. The examples demonstrate an alternative that is applicable to these types of experiments.

EXAMPLE 7.3

In this example, the objective is to compare the effects of two treatments on the proportion of surviving plants when exposed to a certain disease. Suppose that a standard treatment is to be compared to a new experimental treatment, and that experience with the standard treatment suggests that the proportion of plants exposed to the disease that survive averages 15%. It is believed that the experimental treatment can increase that proportion to 25%. The researcher wants to know how many plants per treatment must be observed to have a reasonable chance of detecting such a change.

Some experimental design textbooks have tables giving the needed sample size based on standard formulas for binomial response variables (e.g., Cochran and Cox, 1992). Alternatively, one could use standard power and sample size software, such as PROC POWER in SAS. Either approach yields a required sample size of 250 plants per treatment to have power of 0.80 when a significance level of $\alpha = 0.05$ is used. The GLIMMIX based probability distribution approach would yield the same answer if one uses the program shown in Fig. 7–13. This program assumes a binomial generalized linear model with a logit link. The model is given by

$$\mathrm{logit}(\pi_i) = \beta_0 + T_i$$

FIG. 7–13. GLIMMIX statements to obtain the power for a binomial response for Example 7.3.

```
data binomial;
   input  trt  n  p;
   expected_y = n*p;
   datalines;
  0  250   0.15
  1  250   0.25
run;

proc glimmix data=binomial  initglm;
   class trt;
   model expected_y/n = trt / chisq dist=bin  link=logit;
   ods output  tests3=power_terms;
run;

data power;
   set power_terms;
   alpha = 0.05;
   non_cent_parm = numdf*chisq;
   chi_sq_critical = Cinv(1 – alpha, numdf, 0);
   power = 1 – ProbChi(chi_sq_critical, numdf, non_cent_parm);
run;
```

where π_i is the probability that a plant survives when the ith treatment is applied, β_0 is the intercept and T_i is the ith treatment effect. Note that this model is a true generalized linear model and, hence, uses a χ^2 statistic to test the equality of the π_i. The *chisq* option on the *model* statement requests the χ^2 test. Figure 7–13 also shows the statements needed to compute the power for this model. Note that these statements take into account the fact that the χ^2 distribution is being used as the basis for inference for this model.

Unfortunately, this approach is overly simplistic and misleading for most agronomic research. Most agronomic experiments involve some form of blocking and are often conducted at multiple locations. To see how this affects power, suppose that the proposed experiment is to be performed at four locations. The researcher asks, "If I need 250 plants per treatment, should I divide them equally among the four locations?"

A model that reflects this design is given by

$$\text{logit}(\pi_{ij} \mid L_j, \text{TL}_{ij}) = \beta_0 + T_i + L_j + \text{TL}_{ij}$$

where π_{ij} is the probability that a plant survives when the ith treatment is applied at the jth location, T_i is the ith treatment effect, L_j is the jth location effect, and TL_{ij} is the treatment × location interaction effect. If locations represent a random sample from the target population, then location and treatment × location are random effects, where L_j are independent $N(0, \sigma_L^2)$, TL_{ij} are independent $N(0, \sigma_{\text{TL}}^2)$, and the L_j and TL_{ij} are assumed to be independent.

It is important to understand what the variance components for location and treatment × location signify because they are critical to getting the design correct for this experiment. In categorical data, the ratio $\pi/(1 - \pi)$ represents the odds of the event of interest. The logit of π is the natural logarithm of these odds. The odds ratio is defined to be the odds for the experimental treatment divided by the odds for the reference treatment. The difference between the logits for the two treatments is the log odds ratio. Therefore the variance component σ_L^2 measures the variation in the log odds from location to location averaged over treatments and σ_{TL}^2 measures the variation in the log odds ratio among treatments from location to location. For example, if the probability of a plant surviving averages 0.15 for the reference treatment (and as a result, the log-odds of survival averages −1.73), the actual probability varies from location to location and between treatments over locations. With a little reflection this makes sense because the motivation for multi-site experiments is the implicit assumption that variation exists among locations and one wants to avoid experimental results that are site-specific.

How can one anticipate values of σ_L^2 and σ_{TL}^2 for power or precision analysis and planning experiments? Historical data could provide guidance. Otherwise, the researcher could "guesstimate" the lowest and highest values of π likely to occur among the locations in the population. For example, suppose, based on historical data a researcher "guesstimates" that for the reference treatment $\pi = 0.1$ is the minimum probability of a plant surviving considered plausible at any give location and that $\pi = 0.2$ is the maximum. Converting from the data scale to the model scale, the plausible range of logits across locations is −2.20 to −1.39. The standard deviation can then be approximated as the difference between the maximum and the minimum divided by six, or roughly 0.135. Hence, the variance among logits is approximately $(0.135)^2 = 0.018$. This can serve as an approximation for σ_L^2. If similar variation occurs for the experimental treatment, then odds ratios could vary from 1.0 (when $\pi = 0.2$ for both the reference and experimental treatments) to 3.86 (when $\pi = 0.1$ for the reference treatment and $\pi = 0.3$ for the experimental treatment). The log odds ratio would then vary from 0 to 1.35, yielding a variance of $(0.135/6)^2 = 0.05$ as an approximation for σ_{TL}^2. In this way approximate values for the variances of the location and treatment × location random effects can be obtained.

For this example, round off the approximate variance components obtained above; that is, use "best guesses" of $\sigma_L^2 = 0.02$ and $\sigma_{\mathrm{TL}}^2 = 0.05$, respectively. Suppose the researcher proposes to observe 65 plants per treatment at each of the four locations. Figure 7–14 shows the SAS statements needed to approximate the power using the probability distribution method.

FIG. 7–14. SAS program to determine the approximated power for the multi-location binomial experiment in Example 7.3.

```
data multi_loc_binomial;
   input  trt  n  pi;
   do location = 1 to 4 by 1;
     expected_y = n*pi;
     output;
   end;
   datalines;
 0 65   0.15
 1 65   0.25
run;

proc glimmix data=multi_loc_binomial;
   class  location  trt;
   model  expected_y/n = trt / dist=bin link=logit;
   random  intercept  trt / subject=location;
   parms (0.02) (0.05) / hold=1,2;
   ods  output  tests3=power_terms;
run;

data power;
   set power_terms;
   alpha = 0.05;
   non_cent_parm = numdf*Fvalue;
   F_critical = Finv(1 – alpha, numdf, dendf, 0);
   Power = 1 – ProbF(F_critical, numdf, dendf, non_cent_parm);
run;
```

TABLE 7–5. Approximated and estimated power for 65 plants per location–treatment combination for Example 7.3.

Number of locations	Approximated power†	Estimated power†	Lower confidence limit‡	Upper confidence limit‡
4	0.36	0.277	0.257	0.296
8	0.80	0.838	0.822	0.854

† Approximated power is based on the probability distribution method. Estimated power is based on the simulation method with 2048 simulated samples.

‡ 95% confidence limits for the estimated power.

Since this is a generalized linear mixed model, the test for no treatment effect on the logit scale uses an F statistic. As a result, the subsequent computations necessary to calculate the power are exactly as shown previously in Fig. 7–7. This approach yields a power of 0.36, far less than the power of 0.80 often used in sample size calculations. The reason for the discrepancy is that standard power computations for binomial responses do not account for the variance among locations and, as a result, are vulnerable to dramatically overstating the power and understating the actual sample size requirements. One can vary the number of plants per location by changing *n* and vary the number of locations by changing the *do* statement to examine various design alternatives. With 65 plants per treatment group at each location, we see that eight locations are required to achieve power of at least 0.80 given the assumed variance components (Table 7–5).

Since the response variable is not normally distributed, this is a situation where it is important to use the simulation method to check of the accuracy of the probability distribution method. For various values for the number of sites and the total number of plants, 2048 independent samples were generated according to the model above. Each sample was analyzed using the GLIMMIX model shown in Fig. 7–14 after omitting the *parms* statement. The results obtained were then used to estimate the true power for each combination of the simulation parameters.

For example, both approximated and estimated power values that result by using four and eight sites with 65 plants per treatment group at each site are shown in Table 7–5. Note that the approximated power obtained from the probability distribution method using the GLIMMIX statements in Fig. 7–14 is higher than the power estimate obtained using the simulation method when four locations are used, but that the estimated power obtained using the simulation method is greater than the approximated power obtained from the probability distribution methods when eight locations are used. In other words, the probability distribution method gives a somewhat optimistic power approximation when the experiment is under-powered and a slightly pessimistic approximation when the experiment is adequately powered. Discrepancies aside, both the simulation and probability distribution power analyses give accurate assessments of whether the proposed number of locations is sufficient or not.

Table 7–6 gives results for additional combinations of number of locations and number of plants per treatment per location. Note that the total number of plants

TABLE 7–6. Approximated and estimated power for various numbers of locations and plants per location–treatment combination for Example 7.3.

Number of locations	Plants per location–treatment combination	Total number of plants per treatment	Approximated power†	Estimated power	Lower confidence limit‡	Upper confidence limit‡
10	26	260	0.63	0.622	0.601	0.643
10	43	430	0.80	0.823	0.806	0.839
20	13	260	0.72	0.742	0.723	0.761
20	15	300	0.78	0.771	0.753	0.790
20	16	320	0.80	0.799	0.782	0.817
50	6	300	0.83	0.833	0.817	0.850
132	2	264	0.80	0.811	0.794	0.828

† Approximated power is based on the probability distribution method. Estimated power is based on the simulation method with 2048 simulated samples.

‡ 95% confidence limits for the estimated power.

required decreases as the number of locations increases, but at no point is it possible to obtain 80% power with only 260 plants. Some researchers believe that the algorithm used by GLIMMIX does not produce accurate results when the "cluster size" (i.e., number of plants per location) is small. We observe this to be true for underpowered experiments (e.g., 2 plants per location and few locations), but not when the number of locations is sufficient for adequate power. This underlines the need to design experiments tailored to the distribution of the response variable to be analyzed and not to depend on conventional wisdom. ■

EXAMPLE 7.4

As a variation on Example 7.3 that clearly illustrates the effect of the number of locations on power, suppose there are a total of 600 plants available and that they are to be divided equally between treatments among a number of locations to be used in the experiment. As in the previous example suppose that $\sigma_L^2 = 0.02$ and $\sigma_{\mathrm{TL}}^2 = 0.05$. Using these assumed values for the variance components, what power can be achieved for detecting the difference between the proportions 0.15 and 0.25? Does the power depend on how many locations we use? If so, in what way does it matter?

Table 7–7 shows the power for this test as a function of the number of locations used from 2 to a maximum of 150. There are several things to notice from this analysis. First, we see that across the entire range of the number of locations we could use, the power increases as the number of locations increases. As might be expected, the per-location increase in power is greatest when the number of locations is small. The maximum power attainable is 0.85, which occurs when we use 150 locations. It appears that using between 25 and 30 locations results in a

TABLE 7–7. Approximated and estimated power for 600 total plants for Example 7.4.

Number of locations	Plants per Location	Approximated power†	Estimated power†	Lower confidence limit‡	Upper confidence limit‡	Number of samples†
2	300	0.1302	0.0000	0.0000	0.0000	2048
3	200	0.2640	0.0870	0.0747	0.0992	2047
4	150	0.3887	0.2954	0.2756	0.3151	2045
5	120	0.4818	0.4628	0.4412	0.4844	2044
6	100	0.5486	0.5340	0.5124	0.5556	2043
10	60	0.6844	0.6663	0.6459	0.6868	2041
12	50	0.7170	0.7383	0.7193	0.7574	2037
15	40	0.7485	0.7515	0.7327	0.7703	2024
20	30	0.7787	0.7756	0.7575	0.7938	2028
25	24	0.7960	0.7986	0.7812	0.8161	2031
30	20	0.8072	0.8286	0.8122	0.8451	2025
50	12	0.8288	0.8454	0.8295	0.8612	2011
60	10	0.8340	0.8321	0.8158	0.8484	2025
75	8	0.8391	0.8437	0.8279	0.8595	2028
100	6	0.8441	0.8559	0.8405	0.8712	2019
150	4	0.8491	0.8558	0.8405	0.8711	2018

† Approximated power is based on the probability distribution method. Estimated power is based on the simulation method. 2048 samples were simulated for each number of locations. The number of samples for which the GLIMMIX procedure converged successfully is given in the rightmost column. Section 2.7 briefly discusses the computational issues involved with convergence of the numerical algorithms used.

‡ 95% confidence limits for the estimated power.

power of 0.80. In addition, there is little reason to use more than 30 locations, since beyond this point the per-location increase in power is very low.

This is a situation where it is important to use simulation to verify the power approximations provided by the probability distribution method. Again, 2048 independent samples were generated for each combination of the number of locations and the number of plants per treatment per location. As can be seen in Table 7–7, in several cases the approximated power obtained from the probability distribution method is actually larger than the upper endpoint of the 95% confidence interval estimate of the power obtained from the simulation method. In particular, for this situation the values of the approximated power appear to be too large when fewer than five locations are considered. The power approximations provided by the probability distribution method appear to be most accurate when the number of locations is large, even though in those cases the number of plants at each location is small. This is a somewhat surprising result. The downside is that the proportion of samples for which the GLIMMIX estimation procedure converges tends to decrease as the number of plants per location decreases. ■

7.8 A SPLIT PLOT REVISITED WITH A COUNT AS THE RESPONSE VARIABLE

In Example 7.2 the response variable was assumed to be continuous and normally distributed and the focus of the inference was the treatment × linear rate effect. Generalized linear mixed models were used to evaluate the power profiles of three potential designs (randomized complete block, incomplete block, and split plot) for the experiment. What if the response does not have a normal distribution? The approach presented in that example can be used to evaluate the power profile of one or more designs for other types of responses. In this section we show how this can be accomplished when the response of interest is a count (e.g., number of weeds or insects).

The probability distribution of counts in biological settings has received considerable attention in recent years. Young and Young (1998) provided a good summary of the main issues. Historically the Poisson has been the presumptive distribution for counts. One important characteristic of the Poisson distribution is that the mean and variance of the distribution are equal. This is a very strong assumption, and there is now considerable empirical evidence suggesting that biological count data that satisfy the Poisson assumption are very much the exception (Young and Young, 1998). On the other hand, much evidence supporting the use of other distributions, such as the negative binomial (Section 2.3), has accumulated from field studies over the past several decades.

The motivation for using the negative binomial rather than the Poisson is over-dispersion. Relative to the Poisson distribution, over-dispersion occurs whenever the variance is larger than the mean. It occurs with count data when biological entities (e.g., weeds, insects, mold, viruses) tend to cluster rather than disperse completely at random. The negative binomial distribution can account for events occurring at random with clustering, whereas the Poisson assumes events occurring completely at random. Hence, the negative binomial tends to be a better model for biological counts in many situations, and planning research under the Poisson assumption can result in serious, even disastrous mistakes in assessing sample size requirements.

In this section we focus on the negative binomial distribution. In addition, because count data are often analyzed using a normal approximation with transformations, typically the natural logarithm or square root of the counts, the implications of power analysis from the transformation perspective are also considered.

EXAMPLE 7.5

Generalized linear mixed models for count data typically use the natural logarithm as the link function. For the factorial treatment structure in Example 7.2, when the response is a count that is assumed to have a negative binomial distribution, the conditional model for the split plot design with whole plots in blocks can be written as

$$\log(\mu_{ijk} \mid R_k, w_{ik}) = \mu_{ij} + R_k + w_{ik}$$

where μ_{ijk} is the mean count for the *i*th treatment and *j*th rate in the *k*th block, μ_{ij} is the mean count for the *i*th treatment and *j*th rate, R_k is the *k*th block effect, assumed to be independent $N(0,\ \sigma_R^2)$, w_{ik} is the whole plot error, assumed to be independent $N(0,\ \sigma_W^2)$, and R_k and w_{ik} are assumed to be independent.

The variance component approximations required for a power analysis involving count data can be obtained using an approach similar to that used in Example 7.3 for a binomial response. One begins by determining the variability among counts, from the minimum to the maximum plausible among blocks and among whole plot experimental units for a given treatment × rate combination. Since the generalized linear mixed model models log-counts, we convert the minimum and maximum counts from the data scale (counts) to the model scale (log counts). The range on the log scale divided by six gives an approximation of the standard deviation, which when squared yields the approximate variance. In a split plot this procedure must be used to approximate the block variance as well as the whole plot variance.

If the mean of the negative binomial is denoted by μ then the variance is given by $\mu + k\mu^2$, where *k* is the scale or aggregation parameter ($k = 1/\delta$ in Table 2–2). The scale parameter must be positive. The negative binomial distribution is flexible in that the degree to which the variance exceeds the mean is allowed to vary. In particular, for a fixed value of the mean, the variance varies directly with the value of the aggregation parameter. For values of *k* close to zero, there is little over-dispersion and the variance is close to the mean, as in the Poisson distribution. The over-dispersion increases as *k* increases. In specifying a value of *k* for power and precision analysis using GLIMMIX, one chooses a value of *k* that reasonably approximates the anticipated mean–variance relationship.

One way to obtain a reasonable value of *k* is as follows. Identify the treatment conditions under which the researcher is most familiar with the distribution of counts. For example, in an experiment where an experimental treatment is being compared to a standard treatment, the researcher may be familiar with the distribution of counts under the standard treatment. The researcher can then identify the count that would be expected (μ) under that treatment, as well as the largest and smallest counts that would likely be expected under that treatment. Then an approximate value of *k* can be obtained from

$$k \cong \frac{\left[(\max - \min)/6\right]^2 - \mu}{\mu^2}$$

where max is the largest expected count and min is the smallest expected count under that treatment.

This technique requires the same kind of information regarding the variability of the response as in previous examples and should give a reasonable value of k to use in the calculations.

Suppose that the block variance has been determined to be approximately 0.25 and the whole plot variance approximately 0.15. In addition, the researcher has indicated that when the expected count is 10, then about 50 would be the largest count and 4 the smallest count they would expect to see. With these values, an approximate value for the scale parameter is

$$k \cong \frac{\left[(50-4)/6\right]^2 - 10}{10^2} = 0.49$$

which will be rounded off to $k = 0.5$.

As before, the focus is on inference about the treatment × linear rate effect. Suppose that the researcher is interested in detecting a difference in the linear rate effect when it is three times higher under treatment 2 than it is under treatment 1. In addition, she is interested in determining the number of blocks required to have 80% power of detecting such a difference.

Figure 7–15 shows the SAS statements to create an exemplary data set for this analysis when four blocks are used. The response variable is labeled *expected_count*. Figure 7–16 shows the GLIMMIX statements that provide the values needed to obtain the non-centrality parameter and the degrees of freedom for the power analysis. The *initglm* option on the *proc* statement instructs GLIMMIX to use generalized linear model estimates as initial values for fitting the generalized linear mixed model. The first two terms in the *parms* statement are the block and whole plot variance estimates, respectively. The third term is the aggregation parameter *k*. While Fig. 7–15 and 7–16 are for a split plot design, they can be modified to accommodate other design structures such as the randomized complete block and incomplete

FIG. 7–15. SAS statements to create an exemplary data set for Example 7.5.

```
Data  Split_Plot_with_Counts;
   input trt rate expected_count;
   do block=1 to 4 by 1;
      output;
   end;
   datalines;
  1   1  10
  1   2   9
  1   3   8
  2   1   9
  2   2   6
  2   3   3
run;
```

FIG. 7–16. GLIMMIX statements for the power analysis for negative binomial model for Example 7.5.

```
proc glimmix data=Split_Plot_with_Counts  initglm;
   class block trt rate;
   model  expected_count=trt  rate  trt*rate / dist=NegBin  link=log;
   contrast  'trt x lin rate'  trt*rate  −1  0  1  1  0  −1;
   lsmeans  trt|rate / diff  slicediff=(trt  rate)  ilink;
   random  intercept  trt / subject=block;
   parms  (0.25)(0.15)(0.50) / hold=1, 2, 3;
   ods  output  contrasts=power_terms;
run;
```

TABLE 7–8. Approximated and estimated power for the split plot design with the negative binomial distribution in Example 7.5.

Number of blocks	Approximated power†	Estimated power†	Lower confidence limit‡	Upper confidence limit‡	Number of samples†
4	0.1670	0.1906	0.1662	0.2150	997
10	0.3782	0.4011	0.3689	0.4333	890
20	0.6576	0.6410	0.6100	0.6720	922
27	0.7877	0.8000	0.7744	0.8256	940
28	0.8024	0.7871	0.7611	0.8130	958

† Approximated power is based on the probability distribution method. Estimated power is based on the simulation method. 1024 samples were simulated for each number of blocks. The number of samples for which the GLIMMIX procedure converged successfully is given in the rightmost column. Section 2.7 briefly discusses the computational issues involved with convergence of the numerical algorithms used.

‡ 95% confidence limits for the estimated power.

block alternatives discussed previously. This can be done regardless of the distribution assumed for the counts. As in the case of the normal distribution, the block variance changes depending on the proposed design. That is, if natural blocks of size three are combined into complete but heterogeneous blocks of size six, block to block variability will necessarily decrease as within block (whole plot) variability increases. Increasing within block heterogeneity will also increase over-dispersion.

The power associated with different numbers of blocks can be obtained by varying the upper bound in the *do* statement in Fig. 7–15. The results for various numbers of blocks are given in Table 7–8 for the split plot. With four blocks, there is only approximately a 17% chance of detecting a threefold difference in linear rate effects. To achieve 80% power, 28 blocks would be needed. ■

Two questions arise at this point. First, if we assume a Poisson distribution for the counts, will the results change? If so, how? Second, what if the power analysis is based on a normal approximation using a transformation such as the logarithm or square root of the counts? These questions are considered in the following examples.

EXAMPLE 7.6

This example is a continuation of Example 7.5 for the split plot design in which the response is assumed to follow a Poisson distribution. Since the Poisson generalized linear mixed model is also on the log scale, the process that led us to assuming block and whole plot variances of 0.25 and 0.15, respectively, for the negative binomial would lead us to the same anticipated variance components for the Poisson. However, estimation of the scale parameter *k* in the negative binomial would not be applicable. If one computes the approximated power under the Poisson assumption, a power of 50% for four blocks is obtained. Only eight blocks are needed to obtain a power over 80%. The power for this situation, accounting for over-dispersion using the negative binomial distribution, would only be 31%. Failing to account for over-dispersion by assuming a Poisson distribution generally results in severely underestimating the resources needed for adequate power. ■

EXAMPLE 7.7

This example is a continuation of Example 7.5 for the split plot design in which the transformed counts are assumed to be approximately normally distributed. To assess the power using the methods in Example 7.2 following a transformation such as the logarithm or square root of the count, we would use the same exemplary data set as shown in Fig. 7–15. As with the negative binomial power assessment (Fig. 7–16) we would need to determine the approximate variance components. If the log transformation were used, the variance components for block and whole plot error would be the same as for the generalized linear mixed model with log link. If the square root transformation were used, the variances from the log scale would need to be rescaled to the square root scale. Only the log scale will be considered in detail here. While not shown, the square root transformation produced similar results.

If the normal approximation is used, an estimate of the split plot error variance is required in addition to the block and whole plot variance components. This is where the problem with using the normal approximation to assess power occurs. Assume that as before, the expected smallest count is 4 and the expected largest is 50. Then we could anticipate the split plot variance to be approximately

$$\left[\frac{\log(50)-\log(4)}{6}\right]^2 = 0.177$$

Alternatively, the formula for the variance of the negative binomial could be used to produce an estimate which is then transformed to the log scale. For $k = 0.5$ and $\mu = 10$, the split plot variance of the counts would be $\mu + k\mu^2 = 10 + 0.5(10)^2 = 60$. This is important because it might very well be the variance of counts that appears in literature reviews of similar experiments that are often the source of the variance information in power analyses. Using the delta method (Section 3.2), if the variance on the count scale is 60, the variance on the log scale is given by

$$\left[\frac{\partial \log(\mu)}{\partial \mu}\right]^2 \text{var}(count) = \left(\frac{1}{\mu}\right)^2 \text{var}(count) = \left(\frac{1}{10}\right)^2 60 = 0.60$$

The GLIMMIX statements shown in Fig. 7–17 can be used to assess the power assuming that the estimated split plot variance is 0.177. However, the results will be quite different if the estimated split plot variance is 0.60. Two defensible approaches in this case lead to different variances. Which should be used? There is no clear answer.

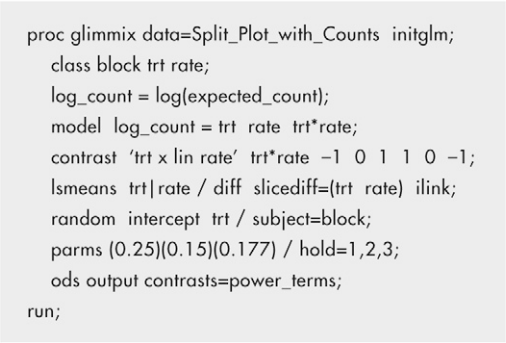
For the normal approximation assuming a split-plot variance of 0.177, the resulting power for four blocks is 48.2% (not shown). For 28 blocks (the required number of blocks assuming the negative binomial), the power is greater than 99.9%. Eight blocks are required to obtain power of at least 80%. This result is similar to what would be obtained with the Poisson distribution. On the other hand, if the power analysis is based on a split-plot variance of 0.60, the power for four blocks is 18.1%, for 28 blocks it is 84.2%, and the required number of blocks for 80% power is 26. All of this assumes that the log counts have an approximately normal distribution.

These results suggest two things. First, using the normal approximation, very different variance estimates and, hence, very different power assessments can be obtained. Using the crude approximation of variance,

$$\left[\frac{\log(\max) - \log(\min)}{6}\right]^2$$

where max is the highest plausible count and min is the lowest plausible count, can result in a very optimistic split plot variance and, hence, a power assessment as misleading as the one based on the Poisson distribution. On the other hand, if the variance from the negative binomial is transformed to the log scale for use

FIG. 7–17. GLIMMIX statements for power analysis for log counts assumed to be approximately normally distributed for Example 7.7.



```
proc glimmix data=Split_Plot_with_Counts  initglm;
   class block trt rate;
   log_count = log(expected_count);
   model  log_count = trt  rate  trt*rate;
   contrast  'trt x lin rate'  trt*rate  -1  0  1  1  0  -1;
   lsmeans  trt|rate / diff  slicediff=(trt  rate)  ilink;
   random  intercept  trt / subject=block;
   parms (0.25)(0.15)(0.177) / hold=1,2,3;
   ods output contrasts=power_terms;
run;
```

with the log count normal approximation, the resulting power calculations are quite different. In this case, they were quite similar to results obtained with the negative binomial, but there is no guarantee what will happen in general.

The second conclusion about the normal approximation follows from the first. There are multiple plausible ways to approximate the variance and, as illustrated above, these include an approach that seems perfectly reasonable but gives a disastrously optimistic assessment of power. Therefore, we suggest avoiding the normal approximation as a tool for power analysis with count data. ■

There are two overriding take-home messages for handling count data. First, do not use the Poisson distribution to plan experiments for count data. There will almost certainly be some level of over-dispersion present in count data. As has been demonstrated above, use of the Poisson distribution can drastically underestimate the number of replications needed. Even in those situations where the amount of over-dispersion is very small and the Poisson may give acceptable results, the negative binomial distribution can still be used since small values of over-dispersion can be accounted for through a small value of the scale parameter *k*. Now that software is available that can fit the negative binomial distribution, we recommend that it be used in planning experiments involving count data. There appears to be no compelling reason to use the Poisson distribution, at least at the planning stages of such an experiment.

Second, planning experiments where count data are the primary responses requires good preliminary information about the anticipated variability. Power computations are sensitive to the scale parameter when using the negative binomial and to the error variance when using the normal approximation. Without good preliminary information, one risks potentially serious errors in planning. One may either overestimate power and hence understate the amount of needed replication or underestimate power and hence overstate the number of replications needed. Obviously, the goal is to avoid either case.

7.9 SUMMARY AND CONCLUSIONS

The generalized linear mixed model based methods to assess power and precision can be applied to any proposed design for which a generalized linear mixed model will be used to analyze the data. The following information is needed:

- the anticipated distribution of the response variable,
- the anticipated magnitude of the variance components implied by the proposed design—for non-normal generalized linear mixed models, care must be taken to express the variance components on the model scale and not the data scale,
- the objectives expressed as testable hypotheses (for power analysis) or interval estimates (for precision analysis),
- for power analysis, the minimum scientifically relevant magnitude of the effect to be detected.

One of the benefits of this approach to power and precision analysis is the requirement that an exemplary data set must be created and GLIMMIX statements to analyze that data set must be written to obtain the needed terms for the power analysis. This is essentially a dress-rehearsal for actual analysis once the data are collected. Subsequently, the researcher is less likely to think, "Now what?" once the data are collected and ready to be analyzed.

Generalized linear mixed model based power or precision analysis also encourages, or should encourage, an early conversation between the researcher and the statistical scientist. As Examples 7.2 and 7.3 clearly illustrate, the terms power analysis and sample size determination often lead researchers to misunderstand the point. Sample size requirements for a badly conceived design can be needlessly high. There are frequently much more efficient designs that researchers cannot be expected to know about, but statistical scientists, given adequate information, can easily suggest. The real question is how to use experimental resources most efficiently, which absolutely mandates involving the statistical scientist in the discussion much earlier than is unfortunately common practice in far too many cases. In an era of tight budgets, this point cannot be emphasized too forcefully.

Finally, the generalized linear mixed model based probability distribution method, in knowledgeable hands, offers a quick way to consider plausible design alternatives. The caveat is that because these methods are relatively new and knowledge about their behavior, especially at the margins, is an active area of research in statistics, the final design choices should be verified via simulation to reduce the chances of unpleasant surprises once the data are collected.

REFERENCES CITED

Cochran, W.G., and G.M. Cox. 1992. Experimental designs. 2nd ed. John Wiley and Sons, New York.

Hahn, G.J. 1984. Experimental design in the complex world. Technometrics 26:19–31. doi:10.2307/1268412

Hinkelmann, K., and O. Kempthorne. 1994. Design and analysis of experiments. Vol. I. Introduction to experimental design. John Wiley and Sons, New York.

Light, R.J., J.D. Singer, and J.B. Willett. 1990. By design: Planning research on higher education. Harvard Univ. Press, Cambridge, MA.

Littell, R.C. 1980. Examples of GLM applications. p. 208–214. *In* Proceedings of the fifth annual SAS Users Group International conference. SAS Institute, Cary, NC.

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

Lohr, V.I., and R.G. O'Brien. 1984. Power analysis for univariate linear models: SAS makes it easy. p. 847–852. *In* Proceedings of the ninth annual SAS Users Group International conference. SAS Institute, Cary, NC.

Mead, R. 1988. The design of experiments: Statistical principles for practical applications. Cambridge Univ. Press, Cambridge, UK.

Milliken, G.A., and D.E. Johnson. 2009. Analysis of messy data. Volume I: Designed experiments. 2nd ed. CRC Press, Boca Raton, FL.

O'Brien, R.G., and V.I. Lohr. 1984. Power analysis for linear models: The time has come. p. 840–846. *In* Proceedings of the ninth annual SAS Users Group International conference. SAS Institute, Cary, NC.

Stroup, W.W. 1999. Mixed model procedures to assess power, precision, and sample size in the design of experiments. p. 15–24. *In* Proceedings of the 1999 Biopharmaceutical Section, American Statistical Association. American Statistical Assoc., Alexendria, VA.

Stroup, W.W. 2002. Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. J. Agric. Biol. Environ. Stat. 7:491–511. doi:10.1198/108571102780

Young, L.J., and J.H. Young. 1998. Statistical ecology: A population perspective. Kluwer Academic Publishers, Norwell, MA.