MIXED MODEL PROCEDURES TO ASSESS POWER, PRECISION, AND SAMPLE SIZE IN THE DESIGN OF EXPERIMENTS

Walter W. Stroup, University of Nebraska
W. W. Stroup, Department of Biometry, University of Nebraska, Lincoln NE 68583-0712

Abstract: Two of the questions most commonly asked of statistical consultants are 1) how should I analyze my data and 2) here's my design - how many replications do I need? These questions are two sides of a common theme, that is, the need for researchers to be as clear as possible about their objectives and the be as familiar as possible with the resources available to them. This paper considers some tools for framing research objectives in terms more amenable to statistical analysis. These, along with relevant aspects of linear mixed model theory, are used to evaluate various design alternatives by estimating the prospective power or precision of competing designs. SAS PROC MIXED is used to illustrate implementation.

1. INTRODUCTION

Viewed narrowly, sample size determination consists merely of deciding how many observations per treatment are needed. However, this assumes that the treatments and the structure of the design of an experiment are known; hence the only remaining question is number of replications. Viewed more broadly, sample size determination ought to be done in the context of a more comprehensive selection of **treatment design** - what treatments are needed to address one's objectives? - and **experiment design** - what is the most efficient way to assign experimental units to treatments?

This paper focuses on using linear mixed model tools, particularly those available in SAS PROC MIXED, to 1) help clarify the treatment design and the comparisons among treatments needed to address one's objectives, 2) help choose from among competing possible experiment designs, and 3) decide what sample size, i.e. number of replications per treatment, is needed.

Nearly every introductory statistics text book introduces the problem of design and sample size. However, most texts confine the introduction to sample size determination to the relatively trivial case of comparing two treatment means in either an independent (completely randomized) or paired (randomized complete block) design. Students are given the formula

$$n \geq 2\left(\frac{\sigma}{\delta}\right)^2 (Z_\alpha + Z_{1-\beta})^2 \qquad (1.1)$$

where n is the minimum required number of experimental units per treatment, σ^2 is the experimental unit variance , δ is the treatment difference to detect, Z_α is the critical value to reject H_0: no treatment difference, and $Z_{1-\beta}$ is the table Z-value corresponding to a desired power of $1-\beta$. Some texts substitute t-values for Z-values. To use (1.1) to determine required sample size, one must know, or have a reasonable idea of

- the minimum treatment difference of interest (δ)
- the magnitude of random variation (σ^2)
- the allowable risk of type 1 error (α-level)
- the desired power (or allowable risk of type 2 error)

However, students are not given any idea what to do if 1) the objectives of the study call for something more complicated than a simple comparison of two treatment means, or 2) if the study requires something more complicated than a completely randomized design (CRD) or randomized complete block design (RCBD). This probably contributes to the overuse of CRD and RCBD even in situations where another design would clearly be more appropriate.

O'Brien and Lohr (1984) presented a method to use ordinary least squares linear model theory to anticipate power. Their method uses SAS PROC GLM to compute the non-centrality parameter of a non-central F under departures from the null hypothesis of no treatment difference deemed to be of interest. The O'Brien-Lohr method applies to any mean comparison that can be tested using an F-statistic: overall equality of treatment means; main effects and interactions in factorial experiments; contrasts among linear combinations of means; or simple pairwise treatment comparisons. However, the method is restricted to designs involving a single source of i.i.d. experimental errors. Hence, it cannot be used to assess repeated measures experiments, split-plot and related designs, or experiments with spatial variation.

This paper presents an extension of the O'Brien-Lohr method to designs whose errors can be assumed approximately multivariate normal but whose variance-covariance structure is of essentially unlimited complexity. As mentioned above, this is particularly useful for planning experiments with repeated measures, split-plot structure, spatial variability, as well as incomplete block designs for which the analysis is to use recovery of inter-block information.

Section 2 presents essential linear model background. Section 3 presents basic programming of SAS PROC MIXED required to implement the method. Section 4 presents examples illustrating the method's various uses. Section 5 presents concluding remarks.

2. LINEAR MODEL BACKGROUND

Consider the linear mixed model

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \mathbf{e} \qquad (2.1)$$

where **y** is a vector of observations, X and Z are matrices of known constants for the fixed and random effects, respectively, $\boldsymbol{\beta}$ is a vector of fixed effect parameters, **u** is a vector of random model effects, and **e** is the error, or residual, vector. Typically, $\boldsymbol{\beta}$ consists of treatment effects, but it may also contain block effects, regression parameters, etc. For **u** and **e**, assume

$$\begin{bmatrix} u \\ e \end{bmatrix} \_ MVN \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}, \qquad (2.2)$$

and, hence, $\mathbf{y} \sim MVN(X\boldsymbol{\beta}, V)$ where $V = ZGZ' + R$. The covariance matrices $G=Var(\mathbf{u})$ and $R=Var(\mathbf{e})$ can have any valid variance-covariance matrix form.

In the mixed model, hypotheses of the form H_0: $K'\boldsymbol{\beta}=0$, where $K'\boldsymbol{\beta}$ is a estimable, can be tested using the generalized F- statistic

$$F = \frac{(K'b)'\,[K'(X'V^{-1}X)^{-1}K]^{-1}(K'b)}{rank(K)} \qquad (2.3)$$

where **b** is the estimate of $\boldsymbol{\beta}$, and V is replaced by its estimate. The generalized F-statistic is distributed approximately $F_{[rank(K).\ \nu,\ \lambda]}$. The denominator degrees of freedom, ν, are the degrees of freedom to estimate $K'(X'V^{-1}X)^{-1}K$. Littell, et. al. (1996) and Kenward and Roger (1997) discuss in detail how ν is determined. The non-centrality parameter, λ, is

$$\lambda = (K'\beta)'\,[K'(X'V^{-1}X)^{-1}K]^{-1}(K'\beta) \qquad (2.4)$$

Under H_0, $\lambda=0$. On the other hand, when H_0 is false, $\lambda>0$. The exact value of λ depends on $K'\boldsymbol{\beta}$, X, and V, that is, the magnitude of departure from H_0, the design and associated replication (sample size), and the variance and covariance components. Power can thus be determined as $P\{F_{[rank(K).\ \nu,\ \lambda]} > F_{crit}\}$, where λ is the value of the non-centrality parameter under the alternative hypothesis of interest, and $F_{crit}=F_{[rank(K).\ \nu,\ 0,\ \alpha]}$, the value of the central F at the designated α-level.

The ordinary least squares F-statistic

$$F = \frac{SS(K'\beta = 0) / rank(K)}{MSE} \qquad (2.5)$$

is a special case of the generalized F in equation (2.3). In (2.5), $SS(K'\beta=0) = (K'b)'[K'(X'X)^{-1}K]^{-1}(K'b)$, and $MSE = (y-Xb)'(y-Xb)/[n-rank(X)]$. The ordinary least squares F is distributed as non-central $F_{[rank(K),\ n-rank(X),\ \lambda]}$, where the non-centrality parameter,

$$\lambda = \frac{(K'\beta)'\,[K'(X'X)^{-1}K]^{-1}(K'\beta)}{\sigma^2} \qquad (2.6)$$

is a special case of (2.4).

For example, consider the linear model for a completely random design with 3 treatments and 3 experimental units per treatment,

$$y_{ij} = \mu + \tau_i + e_{ij}\ ; \qquad i=1,2,3;\ j=1,2,3$$

Suppose one wishes to compare treatment 1 with the average of treatments 2 and 3, i.e.

$$H_0: \tau_1 - \tfrac{1}{2}(\tau_2 + \tau_3).$$

The F-statistic for this test has rank(K)=1 numerator degree of freedom and *N*-rank(X)=9-3=6 denominator degrees of freedom, where *N* is the total number of observations. The non-centrality parameter for the resulting F-statistic is

$$\lambda = n\left[\tau_1 - \left(\frac{\tau_2+\tau_3}{2}\right)\right]^2 oversigma^2 = \frac{n\left[\mu_1 - \left(\frac{\mu_2+\mu_3}{2}\right)\right]^2}{\sigma^2}$$

, where $\mu_i = \mu + \tau_i$.

If $\mu_1= 26$, $\mu_2 = \mu_3= 20$, and $\sigma^2=5$, then

$$\lambda = \frac{3\left[26 - \left(\frac{20+20}{2}\right)\right]^2}{5} = 14.4$$

Then the power of the test of H_0 for the alternative that the null hypothesis is untrue because there is at least a 6 unit difference between μ_1 and the average of μ_2 and μ_3

$$\text{Power} = 1-\beta = P\{F_{(1,6,\lambda=14.4)} > F_{crit}\}$$

where F_{crit} is the critical value of the central F distribution under H_0 for the specified α-level. In this case, $F_{crit} = F_{(1,6,\lambda=0,\alpha)}$. For example, at $\alpha=0.05$, $F_{crit} = F_{(1,6,\lambda=0,0.05)} = 5.99$.The power is thus $P\{F_{(1,6,\lambda=14.4)}>5.99\} = 0.88$. This is the method presented by O'Brien and Lohr (1984).

In the above example given above, X and β follow from the 3-treatment CRD, there are no random model effects, hence no Z or u, and $V=I_9\sigma^2$ where $\sigma^2=5$. The non-centrality parameter, $\lambda=14.4$, can be obtained either from equation (2.6) or the more general (2.4).

However, the mixed model approach is capable of evaluating designs too complex for the O'Brien- Lohr method. For example, consider a split-plot experiment, with a whole-plot factor (A) in a completely randomized design, and a split-plot factor (B). The model for this experiment is

$$y_{ijk} = \mu_{ij} + w_{ik} + e_{ijk}$$

where μ_{ij} is the mean for the treatment combination consisting of the i^{th} level of A and j^{th} level of B, w_{ik} is the whole-plot error effect, assumed i.i.d. $N(0, \sigma_W^2)$, and e_{ijk} is split-plot error, assumed i.i.d. $N(o, \sigma^2)$. This is a mixed model that can be expressed as

$$y = X\beta + Zu + e$$

where β is the vector of μ_{ij}'s, u is the vector of w_{ik}'s, e is the vector of e_{ijk}'s, and X and Z are the design matrices for the treatments and whole-plot experimental units respectively. The variance structure is given by the following:

- $\text{Var}(u)=G=I\sigma_W^2$,
- $\text{Var}(e)=R=I\sigma^2$, and hence
- $V=ZZ'\sigma_w^2+ I\sigma^2$.

Clearly, the non-centrality parameter cannot be evaluated using equation (2.6) but it can easily be obtained from equation (2.4).

Finally, in comparing candidate designs, it is often of interest to assess the anticipated **precision** of competing designs rather than the anticipated **power**. Mead (1988), for example, uses this approach extensively in his *Design of Experiments* text. In mixed model theory, the variance of the estimate of an estimable function, $k'\beta$ is given by

$$\text{Var}(k'\beta) = k'(X'V^{-1}X)^{-1}k \qquad (2.7)$$

Thus, competing designs can be compared for the precision with which they can be expected to estimate functions $k'\beta$ deemed to be of primary interest.

3. COMPUTING POWER AND PRECISION USING SAS *PROC MIXED*

SAS PROC MIXED can compute the non-centrality parameter given in (2.4). This, in conjunction with SAS function statements for the F distribution, can be used to determine power.

PROC MIXED requires the following four steps to compute power:

1. Create a data set with the structure of the design to be assessed. Instead of observed data, use the μ_i's that reflect the departure from H_0 of interest.

2. Run PROC MIXED with the variance and covariance components set at the anticipated values. Use the NOPROFILE and NOITER options (see below) to set the (co)variance components.

3. The MODEL and CONTRAST statements in PROC MIXED compute F values using (2.3). If the anticipated μ_i's as in step 1 and the anticipated (co)variance components as in step 2 are used, these "F values" computed will in fact be the non-centrality parameter λ as in (2.4). Output these values to a new data set.

4. Use SAS function statements for the F distribution to compute power.

Alternatively, you can compute the variance of an estimable function as given in (2.7), allowing assessment of precision. The LSMEANS and ESTIMATE statements compute standard errors, i.e. the square root of $\text{Var}(k'\beta)$ given in (2.7). Using the μ_i's and (co)variance components set in steps 1 and 2, these yield the anticipated standard errors.

For the 3-treatment CRD example from Section 2 above, the SAS code is as follows:

Step 1, create the data set:

```
data a;
 input trt mu @@;
 do rep=1 to 3;
  output;
 end;
cards;
```

Step 2, compute the non-centrality parameters and output them to a new data set:

```
proc mixed noprofile;
 class trt;
 model mu=trt;
 parms (5)/noiter;
 contrast 'chk' trt 2 -1 -1;
 make 'contrast' out=b;
```

In PROC MIXED as set up here, there is no random statement. The only source of random variation is error, and hence the only component of variance is σ^2. The PARMS statement sets the initial value of σ^2 to 5. Ordinarily, MIXED would use this as a starting value for its restricted maximum likelihood (REML) variance estimation algorithm. However, the combined effect of NOPROFILE in the PROC statement and NOITER in the PARMS statement is to prevent the REML algorithm from running and to fix σ^2 at 5. All remaining statistics, specifically the F-values and standard errors of interest, are computed with $\sigma^2 = 5$.

The MAKE statement creates a new data set containing all the computations associated with the CONTRAST statement. Note that the MAKE statement is used in Version 6 of SAS. In version 8, to be released in the near future, MAKE is replaced with ODS statements. Here, the resulting output is:

```
OBS   SOURCE   NDF   DDF     F      P_F

1     chk      1      6    14.40  0.0090
```

Note that the F-value computed here is the same as the non-centrality parameter, λ=14.4, given in section 2.

Step 4, use data set " b" to compute power.

```
data power;  set b;
alpha=0.05;
nc=ndf*f;
fcrit= finv(1-alpha,ndf,ddf,0);
power=1- probf(fcrit,ndf,ddf,nc);
```

FINV and PROBF are SAS functions to compute the value corresponding to a given cumulative probability and the cumulative probability corresponding to a given value, respectively, for the non-central F distribution. See SAS documentation (SAS Institute, 1990) for details. To print the results, use the statements:

```
proc print;
 var source ndf ddf alpha ncparm fcrit power;
```

The resulting output is

```
OBS SOURCE NDF DDF ALPHA  NC  FCRIT   POWER

1   chk    1    6  0.05  14.4 5.987  0.8824
```

Thus, the power of the test for the difference among means given here is 0.8824. You can vary the means or the sample size (number of "reps") in the DATA step or vary the error variance in the PARMS statement to explore various alternatives.

You can also determine the standard errors for various comparisons. For example, you could add

```
lsmeans trt / diff;
estimate '1 vs 2 & 3' trt 2 -1 -1 / divisor=2;
```

to the PROC MIXED statements given above. The DIVISOR option in the ESTIMATE statement causes the coefficients to be divided by 2, hence the estimable function is $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$. The DIFF option in LSMEANS computes the standard errors of all possible pairwise differences among treatment means.

4. SOME EXAMPLES

4.1 Getting Started

To use the mixed model methods presented in sections 2 and 3 effectively, the statistical consultant and client must collaborate in translating experimental objectives into statements that can be either tested or estimated using estimable functions. Usually this means stating objectives in the form of pairwise mean comparisons or contrasts as in the example in section 3.

The following example is simplistic, but is a model for translating objectives into estimable functions. Suppose that a researcher wants to compare an experimental treatment, say a new drug, with an existing standard. To plan the experiment, the word "compare" needs an operating definition. This may be as simple as identifying a response variable and

comparing the mean of the experimental drug to the standard. However, imagine that the researcher says that it's not that simple. The experimental drug is known to have a linearly increasing response to dose

Table 1. Treatment Design[1]

	Low	Medium	High
Standard	Trt 1 (0)	2 (4)	3 (8)
Experi-mental	4 (0)	5 (8)	6 (16)

[1] Numbers in parentheses are the minimum difference from the response at low dose level that is of interest to the researcher.

As a result of this give and take, the researcher also reveals that it is possible that the experimental treatment will increase linearly only up to a point, then plateau, yielding a quadratic regression. Eventually, consultant and client agree that a 2 x 3 treatment design - that is, 2 drugs, experimental and standard, each observed at 3 dose levels, low, medium, and high - is required to provide the desired information. Also, the **primary objective** is to see if the linear response to dose is the same for each drug. Hence, the estimable function that allows the primary objective to be tested is the **linear dose x treatment** contrast.

This give and take of starting with an objective, giving it an operating definition, using it to determine the required treatment design, and finally obtaining a contrast, or set of contrasts, to address the primary objective(s) is a critical aspect of experimental planning. It is an essential prerequisite to determining sample size. Unfortunately, this process is often slighted in introductory statistics, with the result that many researchers (and statisticians as well!) do not understand this process very well. All too often this is reflected in poorly designed experiments or in confusion about how to analyze data once they have been collected.

Once the treatment design and essential estimable functions are selected, one more item of information is needed: what is the minimum difference between the linear dose effects of the two treatments that would be of interest if it in fact existed? Suppose the researcher says that for the standard treatment, the response increases approximately 4 units as the dose increases from low to medium, and another 4 units from medium to high. Suppose the researcher says that the new treatment is of interest if its response to increasing dose level. What is really of interest is whether the linear increase in response to dose is greater, i.e. the slope of the linear regression over dose levels, is greater for the experimental treatment that for the standard.
is at least two times that of the standard. We can summarize the treatment design and minimum treatment effects of interest in Table 1.

4.2 Using Power Analysis to Select a Design and Sample Size

Once the treatment design, estimable functions, and minimum difference of interest are decided, we now can do power analysis. This section illustrates power analysis in the context of selecting a design.

For the example discussed in section 4.1, suppose we have 24 experimental units available, and that the experimental units are in natural subsets of size 4. For example, these might be 24 animals divided into 6 sets, such as litters or weight classes, of 4 animals each. Schematically, the available experimental units can be described as follows

set 1	set 2	set 3	set 4	set 5	set 6

Several designs for assigning experimental units to treatments are possible. We consider the following. .

Design 1 (true PBIB)

blk 1	blk 2	blk 3	blk 4	blk 5	blk 6
1	1	1	4	4	4
2	2	2	5	5	5
3	3	3	6	6	6
4	5	6	1	2	3

Design 2 ("approximate BIB")

blk 1	blk 2	blk	blk 4	blk 5	blk 6
1	1	1	1	2	3
2	2	2	3	4	4

3	5	3	4	5	5
4	6	5	6	6	6

Design 3 (RCBD)

block 1	blk 2	blk 3	blk 4

Note that designs 1 and 2 are incomplete block designs whose blocks are consistent with the natural sets of experimental units described above. Design 3 groups experimental units into complete blocks. This design strategy is common, and not surprising given relative attention randomized complete block designs receive in statistical methods courses relative to incomplete block alternatives. However, design 3 clearly violates the natural variation among the experimental units. Mead (1988) presents a detailed discussion of these design alternatives We now evaluate these designs using the PROC MIXED. First, consider design 1. The data set for design 1 has the block structure given above and the treatments and their means as given in Table 1. For example, the standard treatment, low dose appears in the data set as treatment 1 with response 0. Suppose the variance among the 6 sets is approximately 4 and the variance among experimental units within set is 6. That is, variance among blocks, σ_B^2=4, and variance among experimental units within blocks, σ^2=6.

The PROC MIXED code is

```
proc mixed noprofile;
 class blk trt;

/* do this for intra-block analysis    (blocks fixed) */
 model mu=blk trt;
 parms (6) / noiter;

 /* do this for inter-/intra-block analysis (blocks
random) */
 model mu=trt;
  random blk;
 parms (4) (6) / noiter;

 /* this is the contrast for the primary objective  */
contrast 'trt x lin' trt 1 0 -1 -1 0 1;
 make 'contrast' out=nc;

data pwr;  set nc;
alpha=0.05;
ncparm=ndf*f;
fcrit=finv(1-alpha,ndf,ddf,0);
power=1-probf(fcrit,ndf,ddf,ncparm);
proc print;
 var source ndf ddf ncparm alpha fcrit power;
```

1	5	3	1	5	3
2	6	4	2	6	4
3	1	5	3	1	5
4	2	6	4	2	6

The results:

1. intra-block (fixed block) analysis

```
OBS       SOURCE       NDF    DDF
 1     trt x lin         1      13

  NCPARM    ALPHA     FCRIT      POWER
  10.0000     0.05    4.66719    0.83197
```

2. combined inter- / intra-block (random block) analysis

```
OBS       SOURCE        NDF    DDF
 1      trt x lin         1      13

    NCPARM    ALPHA     FCRIT      POWER
  10.1818     0.05    4.66719    0.83849
```

The difference in power for the fixed versus random block analysis is trivial. More importantly, this design is clearly adequate to address the main objective.

We could add several additional estimate and contrast statements to assess the design for other possible comparisons. For example, consider the following treatment comparisons:

```
estimate '4u df in a' trt 1 -1 0;
estimate '4u df in a' trt 0 1 -1 0;
estimate '8u df in a' trt 1 0 -1;
estimate '4u df uneq a' trt 0 1 0 -1 0;
estimate '4u df uneq a' trt 0 1 0 0 -1 0;
estimate '8u df uneq a' trt 1 0 0 0 -1 0;
estimate '8u df uneq a' trt 0 0 1 0 0 -1;

contrast '4u df in a' trt 1 -1 0;
contrast '4u df in a' trt 0 1 -1 0;
contrast '8u df in a' trt 1 0 -1;
contrast '4u df uneq a' trt 0 1 0 0 -1 0;
contrast '4u df uneq a' trt 0 1 0 -1 0;
contrast '8u df uneq a' trt 1 0 0 0 -1 0;
contrast '8u df uneq a' trt 0 0 1 0 0 -1;
```

The ESTIMATE statements compute the standard error. The CONTRAST statements are used to compute

power. There are several different mean comparisons suspected to yield 4 or 8 unit differences. Because this is a partially balanced design, not all the mean comparisons are estimated with the same precision. Another advantage of using PROC MIXED to assess designs is that one can easily see how "unbalanced" the design really is. Again, the fixed block versus random block results are trivial. The random block results are:

Parameter	Std Error	DF
4u df in a	1.77281052	13
4u df in a	1.77281052	13
8u df in a	1.77281052	13
4u df uneq a	1.82138967	13

The design is adequate to detect 8 unit differences but not 4 unit differences. Even though it is a partially balanced design, the differences in precision among the associate classes are negligible: there is little to be gained by insisting on a balanced design. If 4 unit differences are essential to detect, one would add additional sets of experimental units, being careful to keep the resulting incomplete block design reasonably balanced, until adequate power is achieved. Mead's (1988) approach to setting up common sense incomplete block designs is especially helpful.

How do designs 2 and 3 compare? The same PROC MIXED code given above can be used for design 2. Design 3 requires modification, because its blocking does not correspond to natural variation. For design 3, the variance among blocks is somewhat lower and the variance among experimental units within blocks in somewhat greater, because the blocks include units from 2 sets. Here, the block variance σ_B^2 becomes 2.5 and σ^2 increases to 6.5. The results are:

"Approximate" BIB: $\sigma_B^2 = 4$, $\sigma^2 = 6$

Parameter	Std Error
4u df in a	1.77492984
4u df in a	1.81886522
8u df in a	1.77492984
4u df uneq a	1.82093094
4u df uneq a	1.77492984
8u df uneq a	1.81886522
8u df uneq a	1.81886522

OBS	SOURCE	POWER
1	4u df in a	0.55010
2	4u df in a	0.53019
3	8u df in a	0.98590
4	4u df uneq a	0.55010
5	4u df uneq a	0.52927
6	8u df uneq a	0.98191
7	8u df uneq a	0.98191
8	trt x lin	0.82906

RCBD: $\sigma_B^2 = 2.5$, $\sigma^2 = 6.5$

Parameter	Std Error
4u df in a	1.80277564

Only one standard error is given for the RCBD, because all treatment differences are estimated with equal precision.

OBS	SOURCE	POWER
1	4u df in a	0.54612
7	8u df uneq a	0.98533
8	trt x lin	0.83445

For this variance structure, the three designs are essentially equally good.

What if the variance among sets is larger, so that the violation of the natural sets by the RCBD (design 3) is more severe? To find out, rerun the analysis but change the PARMS statement to reflect a different block variance. For example, suppose the variance among sets is 36, while the within set variance is 6. For the RCBD, the variances change to σ_B^2=30 and σ^2=9. Now the results are:

Design 1: PBIB: $\sigma_B^2 = 36$, $\sigma^2 = 6$

Parameter	Std Error
4u df in a	1.78647400
4u df uneq a	1.85565327

OBS	SOURCE	POWER

```
1    4u df in a     0.54478
3    8u df in a     0.98492
4    4u df uneq a   0.51418
7    8u df uneq a   0.97806
8    trt x lin      0.83294
```

Design 2, Approximate BIB: similar to design 1

Design 3: RCBD: $\sigma_B^2 = 30$, $\sigma^2 = 9$

```
Parameter        Std Error

4u df in a       2.12132034

  OBS       SOURCE      POWER

   1    4u df in a     0.42276
```

Design 1 (split-plot)

A1	A2	A2	A1	A2	A1	A1	A2
1	2	2	1	3	3	1	3
2	1	3	3	1	2	3	2
3	3	1	2	2	1	2	1

A1 and A2 denote the standard and experimental treatment respectively; applied to entire set. 1,2, and 3 denote low, medium, and high dose levels, respectively; applied to individual experimental units within sets.

Design 2 (RCBD)

block 1		block 2		block 3		block 4	
1	2	1	6	5	6	1	2
3	4	2	5	4	3	6	3
5	6	3	4	1	2	5	4

Design 3 (incomplete block design)

blk 1	blk 2	blk 3	blk 4	blk 5	blk 6	blk 7	blk 8
1	1	1	1	2	2	3	4
2	2	3	4	3	4	4	5
3	5	6	5	5	6	6	6

Designs 2 and 3 can be evaluated using the same PROC MIXED code as in section 4.2. The data sets are changed to reflect the different designs. Design 1 is different because of the split-plot structure. The data set must include a variable to identify the whole plot experimental unit (called WPEU below). Also, it is convenient to identify the factors of treatment type (experimental or standard) and dose (low, medium, high) explicitly, rather than combining them as treatments 1 through 6. Assuming the variance among sets is $\sigma_B^2=4$ and within sets is $\sigma^2=6$, the code is

```
proc mixed noprofile;
 class type wpeu dose;
 model mu=type dose type*dose;
 random wpeu(type);
 parms (4) (6) / noiter;
 contrast 'lin x trt' type*dose 1 0 -1 -1 0 1;
 make 'contrast' data=nc;
```

or alternatively you can use the following code, using "treatment" as defined in previous examples:

```
proc mixed noprofile;
 class type wpeu trt;
 model mu=trt;
 random wpeu(type);
 parms (4) (6) / noiter;
 contrast 'lin x trt' trt 1 0 -1 -1 0 1;
 make 'contrast' data=nc;
```

Though the latter code is an unconventional way to analyze a split-plot, it does yield output comparable to the power analyses already run on the competing designs.

How do the designs compare? The following results show power and precision for the "lin x trt"

```
7    8u df uneq a   0.94069
8    trt x lin      0.70307
```

Here, the undesirable effect of blocking contrary to natural variation among the experimental units is obvious. Design 3 is considerably less precise and powerful that designs 1 and 2.

4.3 A More Advanced Example

Consider the same situation described in section 4.1, same treatment design, same objectives. However, suppose instead of 6 sets of 4 units each, there are 8 sets of 3 experimental units each. The following designs are among those that are reasonable to consider:

contrast, as well as all pair-wise mean comparisons. The treatments are coded as 1 to 6 as given in Table 1. The split-plot results were obtained with the alternative PROC MIXED code, given above, to make things consistent. Also, σ_W^2 for the split-plot denotes the whole-plot error variance and is identical to the among sets variance described above. The output:

Split-plot: $\sigma_W^2 = 4$, $\sigma^2=6$

```
CONTRAST        POWER
  trt x lin     0.84991
```

RCBD: $\sigma^2 = 10$ (results from combining sets to form block)

```
 CONTRAST       POWER
  trt x lin     0.65755
```

Approx BIB: $\sigma_B^2 = 4$, $\sigma^2 = 6$

This example involves 16 treatments to be applied in 4 replications to a spatial layout with the 64 experimental units laid out in an 8 x 8 square grid. Spatial correlation is expected among the experimental units. Suppose that it can be modeled with a spherical covariance model with a range of 3, nugget of 0, and sill (error variance) of $\sigma^2=16$.

Two designs are considered:

Design 1 (4 x 4 Lattice - PBIB)

```
 1  2  3  4  1  5  9 13
 5  6  7  8  2  6 10 14
 9 10 11 12  3  7 11 15
13 14 15 16  4  8 12 16
 1  6 11 16  1  8 11 14
 2  7 12 13  2  5 12 15
 3  8  9 14  3  6  9 16
 4  5 10 15  4  7 10 13
```

Design 2 (RCBD *sort of*)

```
 1  2  3  4  5  6  7  8
 9 10 11 12 13 14 15 16
 1  2  3  4  5  6  7  8
 9 10 11 12 13 14 15 16
 1  2  3  4  5  6  7  8
 9 10 11 12 13 14 15 16
```

```
CONTRAST        POWER
 trt x lin      0.82253
```

There is little to choose between the incomplete block and split-plot designs, but the randomized complete block is clearly less desirable. It is important to stress that this type of power analysis and design comparison ***requires*** mixed model methods. For the split-plot one *must* be able to specify a non-trivial variance structure. This ***cannot*** be done with, say, PROC GLM.

4.4 Correlated errors - a spatial example

Another unique capability of the mixed model power analysis described here is the ability to assess correlated errors. The allows power to be assessed for repeated measures designs, for example. Another application is to situations where spatial variation is likely. This section presents an example of the latter.

```
 1  2  3  4  5  6  7  8
 9 10 11 12 13 14 15 16
```

The first design reflects what one would do to use blocking wisely to control local variation without using spatial adjustment. The second is a design that ignores sound design principles assuming that spatial adjustment can recover lost information.

This example addresses a controversy in spatial statistics: critics worry that field researchers may decide that design is unimportant when spatial variation exists, because spatial covariance adjustment can recover the information. This example should lay this myth to rest.

The power analysis uses the following PROC MIXED code. Selected treatment comparisons are computed to show how their precision is affected by their distance from one another in the design.

```
proc mixed noprofile;
 class trt;
 model mu=trt;
 parms (16) (3) /noiter;
 repeated / subject=intercept type=sp(sph)(row col);
 estimate '1 vs 2' trt 1 -1 0;
 estimate '1 vs 4' trt 1 0 0 -1 0;
 estimate '1 vs 5' trt 1 0 0 0 -1 0;
 estimate '1 vs 10' trt 1 0 0 0 0 0 0 0 0 -1 0;
 estimate '1 vs 13' trt 1 0 0 0 0 0 0 0 0 0 0 0 -1 0;
 estimate '1 vs 16' trt 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1;
 contrast '1 vs 2' trt 1 -1 0;
 contrast '1 vs 4' trt 1 0 0 -1 0;
 contrast '1 vs 5' trt 1 0 0 0 -1 0;
```

```
contrast '1 vs 10' trt 1 0 0 0 0 0 0 0 0 -1 0;
contrast '1 vs 13' trt 1 0 0 0 0 0 0 0 0 0 0 0 -1 0;
contrast '1 vs 16' trt 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1;
make 'contrast' out=b;
```

Both designs were evaluated using the above program. Power was calculated assuming treatment 1 is a reference treatment, treatments 2-4 have means 2 units greater than treatment 1, the treatment 5-12 means are 6 units greater than treatment 1, and treatment 13-16 means are 8 units greater than treatment 1. The results:

Design 1

Parameter	Std Error
1 vs 2	1.74566323
1 vs 4	1.96074990
1 vs 5	1.87674747
1 vs 10	2.24386472
1 vs 13	2.13177516
1 vs 16	2.19799419

SOURCE	POWER
1 vs 2	0.15377
1 vs 4	0.09878
1 vs 5	0.48970
1 vs 10	0.78457
1 vs 13	0.73183
1 vs 16	0.73093

The superiority of design 1 is obvious. Lesson: spatial adjustment cannot recover all information. Sound design is essential, despite the bell and whistles in analysis we now have available. **Design does matter**.

5. CONCLUSIONS

There are two main points to be made:

First, mixed model methods, supported by software such as SAS's PROC MIXED and its functions to evaluate probability distributions, allow power and precision analysis on designs of near arbitrary complexity. The methods presented here are relatively easy. But they do not suffer from the limited applicability of other "easy" methods, most of which are based on ordinary least squares, i.e. PROC GLM or something equivalent, which do not handle multiple error terms or correlated errors, at least not without awkward specialized programming.

SOURCE	POWER
1 vs 2	0.20229
1 vs 4	0.17000
1 vs 5	0.87950
1 vs 10	0.74543
1 vs 13	0.95698
1 vs 16	0.94586

Design 2

Parameter	Std Error
1 vs 2	2.10445109
1 vs 4	3.04149811
1 vs 5	3.04018715
1 vs 10	2.13965285
1 vs 13	3.04035997
1 vs 16	3.04358373

Second, power analysis requires statistical consultant and client alike to understand the planning steps that need to occur **prior to** determining sample size. Precise operating definitions of the objectives, leading to precise statements of exactly how each objective is to be addressed statistically, are essential. Sketching an ANOVA, listing sources of variation, or saying "a mean separation test will be used" is not enough. Operating definitions leading clearly and logically to specific estimable functions are *essential*.

The second paragraph cannot be stressed enough. Two elaborations. First, researchers often ask, "how can we supply values for the treatment means to do power analysis? If we knew them, we would not need to do the research!" However, "what are the means" is the wrong question. The right question is "what is the smallest difference of scientific, economic, therapeutic, or whatever criterion applies, importance?" Part of the statistician's job is to get researchers to think in terms of their subject matter. What kind of a difference is relevant, what magnitude makes it important, and how does one recognize it if it occurs?

Second, researchers often ask what to use for variances in power analysis. The answer ought to be "look in previous literature. What variances are typical?" When students in design of experiments classes try do this, they are appalled to learn how few journals publish meaningful measures of variation, at least consistently. Power and precision analysis can

improve the quality of information produced in studies and significantly reduce the cost of information, as well. The potential of power analysis cannot be realized unless journals provide meaningful information about variability as routinely as they do tables of means.

Statisticians can do a real service by acquainting the scientific community with the importance of design, and by making it clear that journals that fail to provide adequate information about variability make design difficult, thereby contributing to cost inflation in research. Every parent who worries about tuition for their children's college education (much of which supports research), every taxpayer who worries about the cost of regulation (and the studies required to satisfy regulatory requirements), and every consumer who pays indirectly for research and development (or the costs of its not being done well) will appreciate our efforts.

6. REFERENCES

Cochran, W.G. and G.M Cox (1957) *Experimental Designs, 2nd edition*. New York: John Wiley and Sons.

Kenward, M.G. and J.H. Roger (1997) "Small sample inference for fixed effects from restricted maximum likelihood." *Biometrics* **53**: 983-997.

Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger (1996) *SAS System for Mixed Models*. Cary, NC: SAS Institute, Inc.

Mead, R. (1988) *The Design of Experiments*. Cambridge, UK: Cambridge University Press.

O'Brien, R.G. and V.I. Lohr (1984) "Power analysis for univariate linear models: the SAS system makes it easy." *SAS Users' Group International: Proceedings of the Ninth Annual Conference*. Cary, NC: SAS Institute, Inc.

SAS Institute (1990) *SAS Language Reference, Version 6, 2nd edition*. Cary, NC: SAS Institute, Inc.

Sample size calculation is usually conducted based on a pre-study power analysis for achieving a desired power for detection of a clinically meaningful dierence at a given level of signicance. In practice, however, sample size required for an intended clinical trial is often obtained using inappropriate test statistic for correct hypotheses, appropriate test statistic for wrong hypotheses, or inappropriate test statistic for wrong hypotheses.Â This book is a useful reference for clinical scientists and biostatisticians in the pharmaceutical industry, regulatory agencies, and academia, and other scientists who are in the related elds of clinical development. [20] discusses power and sample size in the context of the number of repeated runs, when the experiment is intended at uncovering dierences of algorithms on a single problem instance. Their work also suggest a sequential inference procedure, iteratively increasing the sample size and re-testing until an eect is found or a statistical power of 0.8 is reached for a given MRES, but fails to adequately correct for the increase in type-I errors due to multiple hypothesis testing [54, 12].Â It also does not take into account the questions of desired statistical power, sample size calculation, or the denition of a MRES. Bartz-Beielstein [4, 5] discusses the perils of using a sample size that is too large, in terms of the increase in spurious â€œsignicantâ€ results Learn more about power, precision, and sample size in Stata software.Â Say we are planning an experiment to determine whether students who prepare for the SAT exam obtain higher math scores by (1) taking classes rather than (2) studying independently. The national average math score is 520 with a standard deviation of 135. We want to see the power obtained for sample sizes of 100 through 500 when scores increase by 20, 40, 60, and 80 points or, equivalently, when average scores increase to 540, 560, 580, and 600. We type. . power twomeans 520 (540 560 580 600), n(100 200 300 400 500) sd(135) graph. We assumed above that those studying independently would obtain t