STAT 5200 Handout #24: Power Calculation in Mixed Models

Statistical power is the probability of finding an effect (i.e., calling a model term significant), given that the effect is real. ("Effect" here could be main effect, interaction effect, subject effect, block effect, etc.) When considering a proposed experiment, often funding agencies (or institutional review boards) want to see a power calculation to ensure that the proposed sample size will provide sufficient statistical power to justify the experiment. If the power is low, then the experiment will likely be a wasted effort (because even if there is a real effect, you probably won't find it). If the power is excessively high, then the sample size might be wasteful (because you don't need such a high sample size to get adequate power to find the effect of interest).

Several methods are available to perform power calculations.

- For simple statistical tests (1-sample or 2-sample binomial or t-tests, correlation), you can use PROC POWER.
- For basic ANOVA models (factorial, fixed-effects only), you can use PROC GLMPOWER.
- For mixed models, two approaches to consider (both provide approximate power)
    - Simulation method (can be computationally expensive and not flexible, but don't need to know the sampling distribution of the test statistic)
    - Probability Distribution Method (relatively easy to set up, flexible to see how changing simple things [like number of reps] will affect power; BUT you must know the sampling distribution of the test statistic – this is why sampling distribution has been emphasized throughout the semester)

  Both of these are summarized in Gbur et al. (2012) "Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences". ("Generalized" models do not necessarily assume normality, but also allow for other distributions such as Poisson or Negative Binomial – see Handout #26.)

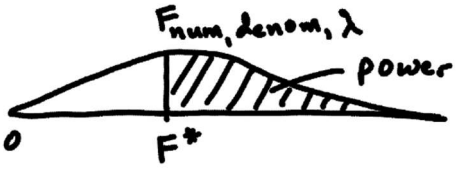
Outline of the Probability Distribution Method: (see example on following page)

1. Create an "exemplary data set" for the proposed design, with data replaced with means (μ_{ijk}'s) representing the minimum difference you would find important to detect (i.e., what difference would you see if the research hypothesis holds).

2. Obtain numerator and denominator DF and λ (non-centrality parameter) for research hypothesis (from "Type III Tests of Fixed Effects" in PROC GLIMMIX output). If H_0 is true, then $F = {MS_{num}}/{MS_{denom}} \sim F_{numDF,denomDF}$. (default $\lambda = 0$)
   If H_0 is false, then $F \sim F_{num,denom}$ with non-centrality parameter λ, where
   $$\lambda = (numDF) \times \left[{EMS_{num}}/{EMS_{denom}} - 1 \right].$$
   Estimate λ by (numDF)×[F statistic from "exemplary data set"]
   – will need cell means ("exemplary" data) and estimates of variance components (for all random effects)

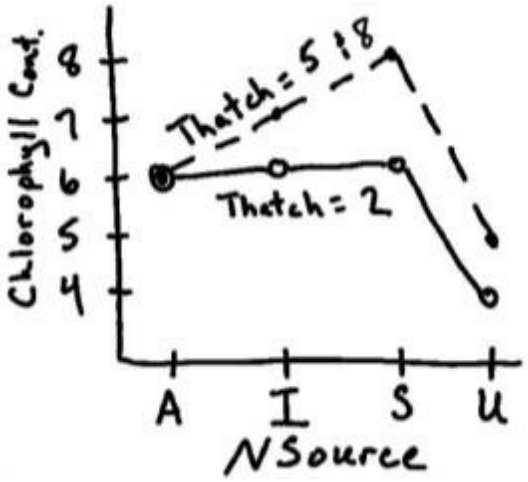
3. Obtain critical value for test statistic (using sampling distribution under H_0) – this is the F^* value from the $F_{numDF,denomDF}$ distribution (with non-centrality parameter $\lambda=0$) that has right-tail area α.

4. Compute power – this is the right-tail area for the F^* value in the $F_{numDF,denomDF}$ distribution (with non-centrality parameter λ as estimated from "exemplary data set").



(Split-Plot)

Example: In the Grass Clippings study (Handout #20, Example 1), the NSource*Thatch interaction was "marginally significant" (p-value .0646). Suppose a researcher believes that the effect of NSource truly does depend on Thatch, and wants to repeat the study to establish this. (Given the 8 years required, this is kind of expensive, but it's just an example.) This researcher hypothesizes that the "true" effects would look like this sketch: (Compare w/ initial interaction plot on p. 4 of Handout #20)



That is, this is their research hypothesis – that the effect of NSource depends on Thatch just like this. Equivalently, this is the minimum effect that they wish to detect. The researcher proposes to use the same levels of NSource and Thatch as before, but this time using four Fields instead of two. What will be their power to detect a significant NSource*Thatch effect (at level $\alpha=0.05$)?

To run the "exemplary data set" through PROC GLIMMIX, we'll need:

- Cell means (μ_{ijk}'s) for fixed effects demonstrating desired effect to be detected. (Here, get these from the hypothesized plot above.)

- Estimates of random effects' variance components – usually based on previous literature (or data from a pilot study) or on "worst-case" scenarios. (Here, get these from analysis of initial data ["Covariance Parameter Estimates" table on p. 3 of Handout #20].)

$$\sigma^2_{Field} = 0.008, \qquad \sigma^2_{NSource*Field} = 0.07, \qquad \sigma^2 = 0.2$$

```
/* Define 'exemplary' dataset
   -- showing minimum effect to be detected
   estY = anticipated cell mean # Chlorophyll Content
 */

data ex;
  input Thatch NSource $ estY;
    do Field=1 to 4 by 1; /* 4 = number of Fields */
      output;
     end;
  cards;
      2   AmmSulph    6
      2   IBDU        6
      2   SCUrea      6
      2   Urea        4

      5   AmmSulph    6
      5   IBDU        7
      5   SCUrea      8
      5   Urea        5

      8   AmmSulph    6
      8   IBDU        7
      8   SCUrea      8
      8   Urea        5
run;

proc print data=ex;
  title1 'Exemplary Set';
run;
```

Exemplary Set

Obs	Thatch	NSource	estY	Field
1	2	AmmSulph	6	1
2	2	AmmSulph	6	2
3	2	AmmSulph	6	3
4	2	AmmSulph	6	4
5	2	IBDU	6	1
6	2	IBDU	6	2
7	2	IBDU	6	3
8	2	IBDU	6	4
…				
41	8	SCUrea	8	1
42	8	SCUrea	8	2
43	8	SCUrea	8	3
44	8	SCUrea	8	4
45	8	Urea	5	1
46	8	Urea	5	2
47	8	Urea	5	3
48	8	Urea	5	4

```
/* Obtain test statistic */
proc glimmix data=ex noprofile;
  class NSource Field Thatch;
  model estY = NSource | Thatch;
  random Field Field*NSource;
  parms (.008)(.07)(.2) / hold=1,2,3;
     /* This sets variance components for random terms,
        in the same order as they appeared in GLIMMIX
        "Covariance Parameter Estimates" table:
        Field  NSource*Field  Residual */
  ods output tests3=power;
  title1 'Dummy Analysis of Exemplary Data Set';
  title2 '(just to get F-statistic)';
run;
```

fit model

fit σ^2_* → all random terms

Dummy Analysis of Exemplary Data Set
(just to get F-statistic)

The GLIMMIX Procedure

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
Field	0.008000	.
NSource*Field	0.07000	.
Residual	0.2000	.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
NSource	3	9	37.94	<.0001
Thatch	2	24	26.67	<.0001
NSource*Thatch	6	24	4.44	0.0037

NOTE: The "ods output" line of code above will send this "Type III Tests of Fixed Effects" table to a data set called "power", which we'll refer to at the top of the following page.

```
/* Approximate power */
data power; set power;
  where Effect = 'NSource*Thatch';
  alpha = 0.05;
  nonCent_param = NumDF*Fvalue;
  FCrit = finv(1-alpha,NumDF,DenDF,0);
  Power = 1 - probf(Fcrit,NumDF,DenDF,nonCent_param);
run;
proc print data=power noobs;
  title1 'Power of test for NSource*Thatch';
run;
```

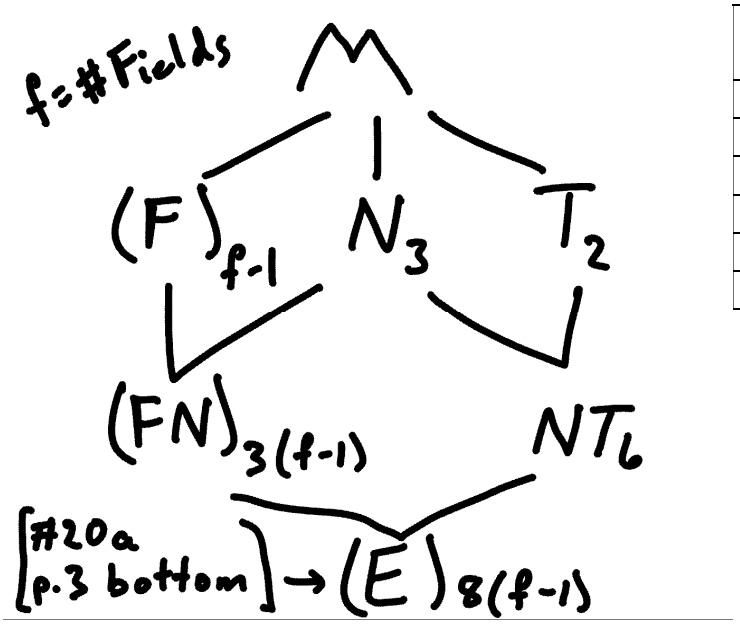
Power of test for NSource*Thatch

Effect	NumDF	DenDF	FValue	ProbF	alpha	nonCent_param	FCrit	Power
NSource*Thatch	6	24	4.44	0.0037	0.05	26.6667	2.50819	0.94753

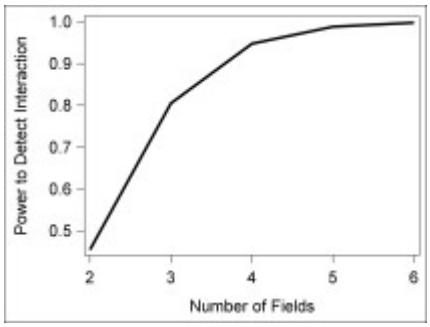
(NOTE: nonCent_param is $\hat{\lambda}$, and FCrit is F^*)

What would power be for different numbers of Fields?

(Just change # Fields in "exemplary" data set in SAS code on p. 3 above)



Number of Fields	Denom. DF	Power
1	0	.
2	8	0.45468
3	16	0.80532
4	24	0.94753
5	32	0.98837
6	40	0.99778



NOTE: The Power increases due to the higher DF for the denominator term in the test statistic. This is where it pays to understand the "error" term for each factor of interest, and how to increase the "error" term DF through true replication (here, more Fields).

A caution … (See Hoenig and Heisey (2001). "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *The American Statistician 55(1):1-6.)*

- Power calculations make perfect sense at the design stage – figuring out necessary sample size, or determining whether the planned sample size will provide power to detect a sufficiently interesting effect size.
- Power calculations make no sense after the experiment (and analysis) are complete. So-called "post-hoc" or "retrospective" power calculations are deeply flawed.