POWER ANALYSIS FOR UNIVARIATE LINEAR MODELS: THE SAS® SYSTEM MAKES IT EASY

Virginia I. Lohr[1], Washington State University
Ralph G. O'Brien[2], The University of Tennessee

Introduction

Power is the probability that a hypothesis test will be statistically significant. In this article we use a hypothetical example to show how easy it is to use PROC GLM and the SAS[3] Supplemental Library functions FPROB and TPROB (Hardison, Quade, and Langston, 1983) to compute power for any hypothesis testable by PROC GLM. (Any other linear models routine, such as PROC REG, can also be used.)

To illustrate this methodology, we will focus upon linear models applied to factorial designs. Associated with factorial designs are many types of hypotheses, including main effects, interactions, subeffects, and simple effects. Each hypothesis leads to a different F-test. The power of a particular F-test (the probability that a particular F-statistic will exceed its critical value) is determined by the specific values for the groups' population means, their common within-group variance, and their sample sizes. Of course, we never know the values for the true means and variances, but we can make reasonable conjectures. By systematically varying the conjectured means, variances, and sample sizes, a thorough "power analysis" can be carried out. Thus researchers can "forecast the productivity" of a proposed design and make adjustments to it before collecting the data. This methodology is perfect for researchers who ask, "how large should my sample size be?" or "what are my best statistical hypotheses for these research hypotheses?"

An Example of Power Analysis

Using a hypothetical example, we have divided the procedure for power analysis for univariate linear models into three main steps: 1) specifying the design, the means, and the hypotheses, 2) computing the power, and 3) evaluating the results.

Step 1: Specifying the Design, the Means, and the Hypotheses

Step 1a: Specifying the Design. Dr. Chen and Dr. Klub are studying the effects of A-toxin and B-toxin on rat liver weight. They believe that increased concentrations of these chemicals in the drinking water of rats will reduce the weight of the rats' livers. A-toxin and B-toxin are believed to act synergistically (interactively), yet they may be nontoxic when present alone.

Chen and Klub propose to conduct a completely randomized, fixed-effects, 3 x 4 factorial design experiment in which the drinking water of individual rats is supplemented with a particular mixed concentration of the toxins. The factor levels and the toxin concentrations they propose to use are:

A-toxin: (1 = 0 PPM, 2 = 10 PPM, 3 = 100 PPM)

B-toxin: (1 = 0 PPM, 2 = 10 PPM, 3 = 100 PPM, 4 = 1000 PPM).

The treatment will last one year, at which time all animals will be sacrificed to obtain the dependent measurement, WEIGHT.

Chen and Klub have sufficient money in their research budget to study as many as 240 rats. They would like to use fewer so that they can have funds for a follow-up study. Chen and Klub therefore have chosen to compare the powers of experiments using N = 120 rats and N = 240 rats (n = 10 and 20 rats per treatment).

Step 1b: Making Educated Guesses About the Population Means and Variances. In order to compute power, Chen and Klub must supply conjectured means for each of their twelve treatments. They must also supply a conjectured error variance. They know from other research that the dependent measure, WEIGHT, has a mean of 100 and a standard deviation near 15 in normal rats. They state that, although the standard deviation should be 15, it could go as high as 20. Based on their knowledge of normal rats and on their belief that the toxins act synergistically, but show no independent effects, Chen and Klub have developed a set of conjectured means (Table 1, Set 1). They have also developed an alternative set of conjectured means which reflect strong independent effects with a small interactive effect (Table 1, Set 2).

Step 1c: Forming Hypotheses to Test. What hypothesis tests are appropriate for this experiment? Chen and Klub want to test for traditional overall main effects and interactions (A, B, A x B). Given the pattern of the conjectured cell means, the consulting statistician suggests that the linear trend contrasts (A-lin, B-lin, A-lin x B-lin) may be tighter questions and therefore more powerful. (They agree that, for a trend analysis, the chosen levels of A-toxin and B-toxin are approximately quantified by an equally spaced metric rather than the raw 0:10:100:1000 metric). The independent toxin effects are assessed by comparing the levels of one toxin within the 0 PPM level of the other toxin (A within b1, B within a1, A-lin within b1, and B-lin within a1). For Set 1, these "simple effects" are null.

Each of the above linear trend contrasts has one degree of freedom. Chen and Klub expect increasing toxicity to lower WEIGHT scores, so it is easy to justify the use of tests with one-tailed rejection regions. They decide to compare the powers of one-tailed and two-tailed tests and choose to use an alpha level of 0.05 for the results.

Table 1. Conjectured Means for the 3 x 4 Designs

		Set 1				Set 2				
B-toxin:		b1	b2	b3	b4		b1	b2	b3	b4
A-toxin:	a1	100	100	100	100	a1	100	99	96	92
	a2	100	100	95	90	a2	99	96	92	86
	a3	100	98	92	84	a3	95	92	86	80

Step 2: Computing the Power

In order to compute power, we must first compute the noncentrality parameter, λ. In the previous paper in these proceedings (O'Brien and Lohr, 1984), we showed that λ measures how much the particular null hypothesis is violated by the population means. We showed that if one has "sample" means that are identical to the conjectured population means, then the sum of squares hypothesis, SSH (in the numerator of the F-statistic), for any testable hypothesis is directly related to the noncentrality parameter,

$$\lambda = SSH/\sigma^2 \qquad (1)$$

where σ^2 is the conjectured population variance. Thus any regular linear models procedure can be used to compute noncentrality parameters: if you can analyze the data for a given design and set of hypothesis tests, then you can also analyze the power of that design.

Step 2a: Computing SSH Values. The SAS code in Table 2 illustrates how easily this can be done. (We have written these commands for readability, not efficiency.) In the DATA SSH step, each cell of the design is entered as one case, with the "data" consisting of the cell codes for the two factors ("A" and "B"), the cell sample size ("N"), and the two conjectured population means ("SET_1_Y" and "SET_2_Y"). In the PROC GLM step, the "FREQ N" statement causes each case to be duplicated N times. Because N can vary from cell to cell, unbalanced designs are easy to specify. The MODEL statement designates SET_1_Y and SET_2_Y as the dependent variables. Thus, for a given set, all 12 cells have "sample" means equaling their conjectured means. The remaining PROC GLM commands are identical to those that would be used for the actual analysis of the data from this experiment: See Freund and Littell (1981) for instruction on these matters.

Table 2. SAS Commands Used to Generate SSH

```
DATA SSH;
     INPUT     A     B      N     SET_1_Y     SET_2_Y;     CARDS;
               1     1     10       100         100
               1     2     10       100          99
               1     3     10       100          96
               1     4     10       100          92
               2     1     10       100          99
               2     2     10       100          96
               2     3     10        95          92
               2     4     10        90          86
               3     1     10       100          95
               3     2     10        98          92
               3     3     10        92          86
               3     4     10        84          80

PROC GLM; CLASS A B; FREQ N;
 TITLE ANALYSIS OF THE 3X4 DESIGN WITH N = 10;
 MODEL SET_1_Y SET_2_Y = A B A*B; MEANS A*B;
  CONTRAST 'A LINEAR'  A 1 0 -1;
  CONTRAST 'B LINEAR'  B 3 1 -1 -3;
  CONTRAST 'A LINEAR X B LINEAR'  A*B 3 1 -1 -3 0 0 0 0 -3 -1 1 3;
  CONTRAST 'A WITHIN B1' A 1 -1 0 A*B 1 0 0 0 -1 0 0 0 0 0 0 0,
                         A 1 0 -1 A*B 1 0 0 0 0 0 0 -1 0 0 0;
  CONTRAST 'B WITHIN A1' B 1 -1 0 0 A*B 1 -1 0 0 0 0 0 0 0 0 0,
                         B 1 0 -1 0 A*B 1 0 -1 0 0 0 0 0 0 0 0,
                         B 1 0 0 -1 A*B 1 0 0 -1 0 0 0 0 0 0 0;
  CONTRAST 'A LINEAR WITHIN B1' A 1 0 -1 A*B 1 0 0 0 0 0 0 -1 0 0 0;
  CONTRAST 'B LIN WITHIN A1' B 3 1 -1 -3 A*B 3 1 -1 -3 0 0 0 0 0 0 0 0;
```

A sample page from the resulting printout, containing the SSH for the means from Set 1 with 10 rats per treatment, is shown in Table 3. The sum of squares for error equals zero, but SAS does not balk at this. No F values can be given, but PROC GLM still lists the SSH. Under SOURCE, the degrees of freedom (DF) and the hypothesis sums of squares (SS) for A, B, and A*B are listed. The DF and SS associated with each specific CONTRAST are given following the SOURCE. The results from the MEANS statement (not shown) are simply used to verify that the sample means equaled the conjectured means.

Table 3. Results of PROC GLM Commands Used to Calculate SSH

ANALYSIS OF THE 3 X 4 DESIGN WITH N = 10
GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: SET_1_Y
FREQUENCY: N

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	11	3089.16666667	280.83333333
ERROR	108	0.00000000	0.00000000
CORRECTED TOTAL	119	3089.16666667	

SOURCE	DF	TYPE III SS	F VALUE	PR > F
A	2	851.16666667	.	.
B	3	1429.16666667	.	.
A*B	6	808.33333333	.	.

CONTRAST	DF	SS	F VALUE	PR > F
A LINEAR	1	845.00000000	.	.
B LINEAR	1	1320.16666667	.	.
A LINEAR X B LINEAR	1	729.00000000	.	.
A LINEAR WITHIN B1	1	0.00000000	.	.
B LINEAR WITHIN A1	1	0.00000000	.	.
A WITHIN B1	2	0.00000000	.	.
B WITHIN A1	3	0.00000000	.	.

Step 2b: Computing the Power for Each Proposed Test. Once the SSH values have been computed, these values, along with their degrees of freedom, are entered as the data for a short SAS routine which calculates the value for the power of each test by using the statistical functions FINV and FPROB (see Hardison et al.).

A listing of the SAS commands used to compute power for the chosen tests for both sets of conjectured standard deviations is shown in Table 4. Titles to identify each hypothesis ("TITLE"), the hypothesis sums of squares ("SSH"), the degrees of freedom associated with each hypothesis ("DFH"), the error degrees of freedom ("DFE"), and the significance levels ("ALPHA") are entered in the DATA FPOWER step. The critical value ("FCRIT") for the null distribution is determined with the FINV function, using zero as the noncentrality parameter. "FNC15" is the noncentrality parameter for $\sigma = 15$ and is used in the FPROB function to calculate "FPW15", the power for the two-tailed test when $\sigma = 15$. Similar statements are used to obtain the power when $\sigma = 20$. The PROC PRINT command generates a printout of the powers for the two-tailed tests.

To calculate the power for the one-tailed test, a new data set ("TPOWER") is created, consisting of all single degree of freedom tests from the existing data set ("FPOWER") (Table 4). A new critical value ("TCRIT") is created using TINV, with zero as the noncentrality parameter. The noncentrality parameter for a one-tailed test is equal to the square root of the noncentrality parameter for a two-tailed test; these ("TNC15" and "TNC20") are calculated in this example by dividing the square root of SSH by σ. The function, TPROB, is used to calculate "TPW15" and "TPW20", the powers for the one-tailed tests. Again, PROC PRINT generates a printout of the powers.

Table 4. Listing of Program Used to Generate Power

```
DATA FPOWER;
   INPUT TITLE & $28.  SSH  DFH  DFE  ALPHA;
     FCRIT=FINV(1-ALPHA,DFH,DFE,0);
     FNC15=SSH/(15**2);
          *NONCENTRALITY PARAMETER FOR SIGMA=15;
     FPW15=1-FPROB(FCRIT,DFH,DFE,FNC15);
          *THIS FIGURES POWER FOR 2-TAILED TEST FOR SIGMA=15;
     FNC20=SSH/(20**2);
          *NONCENTRALITY PARAMETER FOR SIGMA=20;
     FPW20=1-FPROB(FCRIT,DFH,DFE,FNC20);
          *THIS FIGURES POWER FOR 2-TAILED TEST FOR SIGMA=20;

   CARDS;
     A MAIN SET 1 N=10          851.667     2     108     .05
     B MAIN SET 1 N=10         1429.167     3     108     .05
     A X B SET 1 N=10           808.333     6     108     .05
     A LINEAR SET 1 N=10        845.000     1     108     .05
     B LINEAR SET 1 N=10       1320.167     1     108     .05
     .
     .
     .
  PROC PRINT;  ID TITLE;  VAR FPW15 FPW20;
     TITLE POWER OF THE TWO-TAILED TESTS;
DATA TPOWER;  SET FPOWER;  IF DFH=1;
     TCRIT=TINV(1-ALPHA,DFE,0);
     TNC15=(SSH**0.5)/15;     TPW15=1-TPROB(TCRIT,DFE,TNC15);
     TNC20=(SSH**0.5)/20;     TPW20=1-TPROB(TCRIT,DFE,TNC20);
  PROC PRINT;  ID TITLE;  VAR TPW15 TPW20;
     TITLE POWER OF THE ONE-TAILED TESTS;
```

Step 3: Evaluating the Results

Table 5 gives the $\alpha = .05$ power of each of the proposed cases of the 3 x 4 design. These results force Chen and Klub to draw some hard conclusions about their proposed study. If the population means actually resemble Set 1, as they suspect, there is a discomforting probability that no synergistic effects will be found: the most powerful interaction test ($\sigma = 15$, $n = 20$, A-lin x B-lin, one-tailed) has a power of .81. That means that Chen and Klub have a chance of only 81% of correctly rejecting that null hypothesis ($\alpha = .05$) with this proposed design. If fewer rats are used or if σ approaches 20, there is a considerable drop in power. The power of the traditional A x B test is unacceptably low.

If the population means resemble Set 2, it is unlikely that any interaction test will be significant: the most powerful interaction test drops to a power of .30. The main effects are reasonably powerful, thus data under Set 2 are likely to indicate (correctly) that independent toxin effects exist. Unfortunately, however, this conclusion is not likely to be supported by the purest tests of independent effects: even the best WITHIN bl and WITHIN al tests have powers of only .28 and .56.

Table 5. A Power Analysis of the 3 x 4 Design (α = .05)

			Set 1				Set 2			
			σ = 15		σ = 20		σ = 15		σ = 20	
	df	n:	10	20	10	20	10	20	10	20
A	2		.39	.69	.23	.43	.61	.90	.37	.67
B	3		.53	.86	.31	.59	.79	.98	.52	.85
AxB	6		.23	.47	.14	.27	.08	.11	.06	.08
A-lin	1		.48	.78	.30	.53	.71	.95	.47	.76
A-lin (1-tail)	1		.61	.86	.42	.66	.81	.97	.60	.85
B-lin	1		.67	.93	.44	.73	.90	.99	.69	.94
B-lin (1-tail)	1		.78	.96	.56	.82	.95	.99	.79	.97
A-lin x B-lin	1		.43	.72	.27	.48	.12	.20	.09	.13
A-lin x B-lin (1-tail)	1		.56	.81	.38	.60	.20	.30	.15	.21
A w/ b1	2		.05	.05	.05	.05	.10	.15	.08	.11
B w/ a1	3		.05	.05	.05	.05	.16	.31	.11	.18
A-lin w/ b1	1		.05	.05	.05	.05	.11	.18	.09	.12
A-lin w/ b1 (1-tail)	1		.05	.05	.05	.05	.18	.28	.14	.20
B-lin w/ a1	1		.05	.05	.05	.05	.24	.43	.16	.27
B-lin w/ a1 (1-tail)	1		.05	.05	.05	.05	.35	.56	.24	.38

Is There a Better Design?

Chen and Klub conclude that if they run the 3 x 4 design, they must use all 240 rats. Even so, they may still get inconclusive results, especially if the population means resemble Set 1 and σ approaches 20. They are not satisfied with this, so they decide to investigate a "bare-bones" 2 x 2 design, using the corner cells of the 3 x 4 design (Table 1). This allows them to contrast the most extreme groups and to triple the cell sizes.

The powers for the 2 x 2 design (Table 6) are considerably better. For Set 1, the A, B, and A x B tests have the same power. Even if σ = 20, one-tailed powers of .93 can be achieved using all 240 rats. If only 120 rats (n = 30) are used, the power for the one-tailed tests of A, B, and A x B lie between .70 (for σ = 20) and .90 (for σ = 15).

Table 6

A Power Analysis of the 2 x 2 Design (α = .05)

	σ = 15		σ = 20	
n:	30	60	30	60
Set 1				
A, B, AxB	.83	.98	.58	.87
A, B, AxB (1-tail)	.90	.99	.70	.93
Set 2				
A	.87	.99	.64	.91
A (1-tail)	.92	.99	.75	.95
B	.99	.99	.88	.99
B (1-tail)	.99	.99	.93	.99
AxB	.24	.44	.16	.27
AxB (1-tail)	.35	.56	.24	.38
A w/ b1	.25	.44	.16	.28
A w/ b1 (1-tail)	.36	.57	.25	.39
B w/ a1	.54	.83	.34	.59
B w/ a1 (1-tail)	.66	.90	.46	.70

The powers under the Set 2 conjectured means demonstrate that the A and B effects are very likely to be found significant at the .05 level, even with n = 30 animals per cell. The powers of the much weaker A x B test are improved, but not outstanding: the maximum is .56. The simple effects (e.g. A WITHIN b1) are considerably more powerful than their linear trend counterparts of the 3 x 4 design, although the A WITHIN b1 tests are still weaker than Chen and Klub might desire.

What design should the consulting statistician recommend to Dr. Chen and Dr. Klub? None. It is their responsibility to assess the various trade-offs and come up with some compromise solution. The 2 x 2 design is appealing, but will its conclusions be generalizable enough to effectively answer the questions which underlie the research? If the 2 x 2 design is adopted, should they gamble by using n = 30, so that funds are available for another small study? Should they investigate a third design, such as a 3 x 3? These are questions for Chen and Klub. The power analyses provide the information to make informed decisions about these matters.

Finding Sample Sizes for Specified Powers. Another approach which Chen and Kluh could have taken is to specify the desired level of power and ask how large the sample size would need to be to obtain that level of power. An algorithm to accomplish this was discussed in O'Brien and Lohr (1984); a short SAS routine to implement it is contained in the appendix below. We tested 20 situations, all of which converged satisfactorily in 3 iterations or fewer.

Conclusions

Every researcher knows that more careful planning can only improve the probability that research will produce meaningful results. Power analyses should be part of that planning. Researchers can readily perform their own power analyses using the same kind of consulting support that statisticians regularly provide them

for data analyses, because the calculations can be performed by one's favorite software for general linear models. The methodology advocated here is applicable to any test that can be computed with any software that performs one of the special cases of the univariate general linear (regression) model. This includes ANOVA designs with blocking factors, nested factors, unbalanced cell sizes, missing cells, and covariates. These methods coupled with the SAS Supplemental Library functions FPROB and TPROB now make the computation of power simple and quick.

Footnotes

[1]Supported by a Hilton A. Smith Graduate Fellowship from The University of Tennessee.

[2]Supported by a Faculty Development Award sponsored by The Park National Bank, Knoxville, Tennessee. Requests for reprints should be addressed to Ralph O'Brien, Statistics Department, Stokely Management Center, The University of Tennessee, Knoxville, TN 37996-0532.

[3]SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA.

References

Freund, R. J. and Littell, R. C. (1981), SAS for Linear Models, Cary, N. C.: SAS Institute Inc.

Graybill, F. A. (1976), Theory and Applications of the Linear Model, North Scituate, Mass.: Duxbury Press.

Hardison, D., Quade, D., and Langston, R. D. (1983), "Nine Functions for Probability Distributions," in SUGI Supplemental Library Guide 1983 Edition, Cary, N. C.: SAS Institute Inc., 229-236.

Johnson, N. L. and Kotz, S. (1970), Continuous Univariate Distributions-2, Boston: Houghton Miffin.

O'Brien, R. G. and Lohr, V. I. (1984), "Power Analysis for Linear Models: The Time Has Come," in Proceedings of the Ninth Annual SAS Users Group International Conference, Cary, N. C.: SAS Institute Inc.

Rao, C. R. (1973), Statistical Inference and Its Applications (2nd Ed.), New York: John Wiley.

Searle, S. R. (1971), Linear Models, New York: John Wiley.

Appendix

SAS commands used to program a simple algorithm to determine the per-cell sample size, n, required for nominal power, p.

```
DATA FINDN;
NEW: INPUT NAME & $16. NGROUPS DFH SSH1 SIGMA
       ALPHA POWER;
     LAMBDA1=SSH1/SIGMA**2;
     OLDN=10;
     I=0;
ITERATE:
     I=I+1;
     DFE=NGROUPS*(OLDN - 1);
     FCRIT=FINV(1 - ALPHA,DFH,DFE,0);
     NONCENT=FNONCT(FCRIT,DFH,DFE,1 - POWER);
     NEWN=NONCENT/LAMBDA1;
     If ((OLDN - NEWN)**2 LT .01) THEN GO TO
       STOPITER;
     OLDN=NEWN;
     IF (I GT 20) THEN DO;
        PUT 'DID NOT CONVERGE';
        GO TO STOPITER;
        END;
     GO TO ITERATE;
STOPITER:
     NPERCELL=ROUND(NEWN+.25);
     TOTAL_N=NPERCELL*NGROUPS;
     DFE=TOTAL_N - NGROUPS;
     FCRIT=FINV(1 - ALPHA,DFH,DFE,0);
     NONCENT=NPERCELL*LAMBDA1;
     TRUPOWER=1 - FPROB(FCRIT,DFH,DFE,NONCENT);
     OUTPUT;
CARDS;
A MAIN EFFECT  12 2 1460.00 15 .05 .70
A MAIN EFFECT  12 2 1460.00 20 .05 .70
A MAIN EFFECT  12 2 1460.00 15 .05 .90
PROC PRINT;
     ID NAME;
     VAR NGROUPS DFH SSH1 SIGMA ALPHA NPERCELL
       TOTAL_N TRUPOWER;
```