TEACHING POWER ANALYSIS USING REGULAR STATISTICAL SOFTWARE

Ralph G. O'Brien
University of Tennessee

1. Introduction

Until a few years ago I used to feel guilty about how I taught statistical planning, specifically, power analysis and sample-size choice. Of course I did give the typical introductory lectures and hand-workable problems on this topic: We computed the power and determined prudent sample sizes for the binomial test of H_0: $\pi = \pi_0$ versus H_A: $\pi > \pi_0$ when $\pi = \pi_A > \pi_0$ is true. And we devoted similar efforts to normal theory means testing in the one-sample and two-sample situations, pretending that the variance is known. But as an applied statistics sequence progresses to computer-based data analyses involving realistic applications of general linear models and log-linear models, the methods covered are determined in large part by the capabilities of the software being used. The major statistical packages have no general tools for statistical planning, so for years I avoided teaching "realistic" power analysis.

Not any more. This paper presents a simple, unifying approach for teaching and computing power analysis for research designs involving linear and log-linear modeling. The principal advantage of this approach is that it borrows concepts, terminology, and software commonly used for data analysis within these systems. The scheme allows one to easily study the power of any test that can be performed with one's favorite linear or log-linear models routine, thereby making the method more flexible, precise, and "friendly"

than table-based methods, such as those by Cohen (1977) and others.

2. The General Linear Model

One common formulation of the normal-theory general linear hypothesis test is as follows. All design matrices, X, are full column rank. Denote the true (in practice, unknown) model for N observations as $\mathbf{y} = \mathbf{X}_T\beta_T + \mathbf{e}$. Consider a "reduced" (null) design, $\mathbf{X}_R$, and a "fuller" (alternative) design, $\mathbf{X}_F$. Without loss of generality, we take $\mathbf{X}_R$ nested within $\mathbf{X}_F$, which, in turn, is either nested within $\mathbf{X}_T$ or identical to it. The ranks of $\mathbf{X}_T \supseteq \mathbf{X}_F \supset \mathbf{X}_R$ are $r_T \geq r_F > r_R$. The ordinary least squares estimates are $\mathbf{b}_F = (\mathbf{X}_F'\mathbf{X}_F)^{-1}\mathbf{X}_F'\mathbf{y}$, and $\mathbf{b}_R = (\mathbf{X}_R'\mathbf{X}_R)^{-1}\mathbf{X}_R'\mathbf{y}$. Testing the null model versus the alternative model uses

$$F = \frac{SSH(\mathbf{y}, \mathbf{X}_F, \mathbf{X}_R)/(r_F - r_R)}{SSE(\mathbf{y}, \mathbf{X}_T)/(N - r_T)} ,$$

where the sum of squares hypothesis is

$$\begin{aligned} SSH(\mathbf{y}, \mathbf{X}_F, \mathbf{X}_R) &= SSE(\mathbf{y}, \mathbf{X}_R) - SSE(\mathbf{y}, \mathbf{X}_F) \\ &= (\mathbf{y} - \mathbf{X}_R\mathbf{b}_R)'(\mathbf{y} - \mathbf{X}_R\mathbf{b}_R) - (\mathbf{y} - \mathbf{X}_F\mathbf{b}_F)'(\mathbf{y} - \mathbf{X}_F\mathbf{b}_F) \\ &= \mathbf{y}'\mathbf{X}_F\mathbf{b}_F - \mathbf{y}'\mathbf{X}_R\mathbf{b}_R . \end{aligned} \tag{2.1}$$

$\mathbf{X}_T$ is either assumed to be $\mathbf{X}_F$ or is taken to be the most complete model possible, such as when we use the within-cells error term in analysis of variance (ANOVA) testing. Under the usual assumption that the true residuals are independent $N(0, \sigma^2)$ variates, F is distributed as an F random variable with $(r_F - r_R)$ and $(N - r_T)$ degrees of freedom and *noncentrality*

$$\lambda = SSH(\mathbf{X}_T\beta_T, \mathbf{X}_F, \mathbf{X}_R)/\sigma^2 = (\beta_T'\mathbf{X}_T'\mathbf{X}_F\beta_F - \beta_T'\mathbf{X}_T'\mathbf{X}_R\beta_R)/\sigma^2, \tag{2.2}$$

where $\beta_F = (\mathbf{X}'_F\mathbf{X}_F)^{-1}\mathbf{X}'_F\mathbf{X}_T\beta_T$ and $\beta_R = (\mathbf{X}'_R\mathbf{X}_R)^{-1}\mathbf{X}'_R\mathbf{X}_T\beta_T$ are the expected values of $\mathbf{b}_F$ and $\mathbf{b}_R$ under the true model.

Equations (2.1) and (2.2) may be technically explicit, but they hide the simple relationship between the familiar F statistic and the unfamiliar noncentrality parameter. It is much easier to comprehend and remember

$$F = \frac{SSH(sample)/DFH}{s^2} ; \quad \lambda = \frac{SSH(population)}{\sigma^2} ,$$

where DFH is the degrees of freedom for the hypothesis. Any specialized F statistic already familiar to students (e.g. Fs for testing one or more regression predictors, the overall test in a one-way design, preplanned contrasts, main effects, etc.) can be expressed in this way. The main point to teach is that λ is simply DFH times the population analog of the F statistic. If there is no difference between the ways that the null and alternative models fit the expected value of y, then $\lambda = 0$, and F has a *central F distribution*, which is tabled and programmed everywhere so that we can get "p values." If the alternative model is a better representation of the population, then $\lambda > 0$, and F is said to have a *noncentral F distribution*. Tables and charts for noncentral F distributions are also available, but using them is tedious and prone to human error. Give this work to computers!

But even today's computer-using students should still master major concepts such as power. Point out to them that whereas the Type I error rate is the probability that a random F variate will exceed the critical value, F_{crit}, given $\lambda = 0$, power is also the probability that F exceeds F_{crit}, albeit given some particular $\lambda > 0$. A central and a noncentral F distribution are illustrated in Figure 1. Note that when the plots are separated and aligned in this way, rather than overlapping them as is commonly done in statistics texts, they more clearly show the parallelism between Type I error rate and power. When students are given the fact that

$$E[F] = (\lambda/DFH + 1)(N - r_T) / (N - r_T - 2) \approx \lambda/DFH + 1,$$

they see that increasing λ shifts the distribution upward, resulting in increased power.

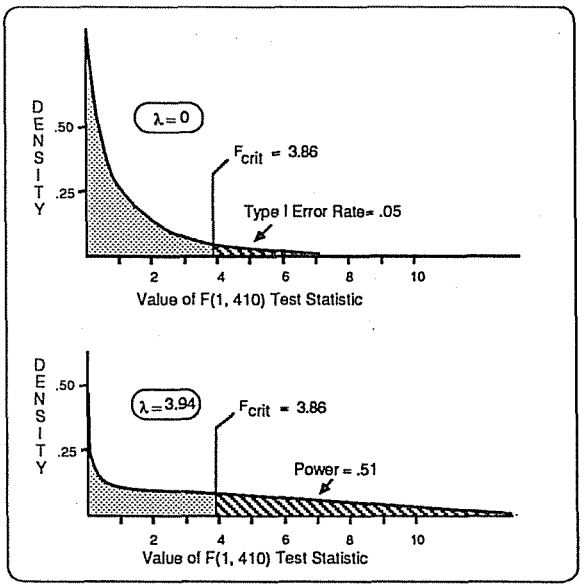

Figure 1: A Central and a Noncentral F Distribution

Noncentrality is directly affected by the sample size. Let $\lambda(N)$ be the noncentrality based on a sample size of N, e.g. equation (2.2). It is easy to show that if the sample size is uniformly changed by a multiplicative factor, m, then the new noncentrality is

$$\lambda(m \cdot N) = m \cdot \lambda(N) . \qquad (2.3)$$

Applied to ANOVA for example, (2.3) implies that if all of the cells' sample sizes are doubled, then the noncentrality is doubled.

To get one's regular linear models software to compute λ, first construct an artificial dependent variable, $\mathbf{y}^* = \mathbf{X}_T\beta_T$, and use it just as if it were a regular dependent variable. All computed SSH(sample) "statistics" will be SSH(population) values computed for a *base total sample size*, N. Noncentrality values for various total sample sizes and standard deviations can be formed using $\lambda(m \cdot N, \sigma^2) = m \cdot SSH(population)/\sigma^2$.

For one-tailed tests with DFH = 1, we use the noncentral t distribution to compute power. The t statistic and its noncentrality parameter are

$$t = F^{1/2} = SSH(sample)^{1/2}/s ; \qquad \delta = \lambda^{1/2} = SSH(population)^{1/2}/\sigma .$$

3. Example

Examples always teach best. This one is hypothetical and uses SPSS^X; O'Brien (1986a) applies SAS to the same problem. My purpose here is to illustrate the three phases of a power analysis: 1) defining the research design, specifying scenarios for the population parameter values (here, for the means and the within-cell standard deviation), and developing a set of statistical hypotheses that conform to the research questions; 2) computing and tabling the power probabilities; and 3) evaluating those power probabilities.

3.1 Design, Scenarios, and Hypotheses

Dr. Cathy Terr, a cardiologist, is planning to study treatments designed to help men who need to lower their blood LDL cholesterol levels. Each man is to be randomly assigned to one cell of the 2 x 3 design formed by crossing the two-level diet and exercise (DIEX) factor (whether or not the man is "required" to participate in the "Healthy Heart Support Group") with the three-level DRUG factor (placebo, drug A, and drug B). Drugs A and B are competing versions of the same basic preparation. The dependent measure is the percentage change in LDL from pretest to four months after onset of treatment.

When Dr. Terr consults her statistician for guidance on appropriate sample sizes, the statistician *patiently* engages her in a dialog to obtain reasonable conjectures about the

unknown population values. Two scenarios for the set of population means are specified, which are listed within the SPSS^X statements below. Two conjectures for the common within-cell standard deviation are also specified: $\sigma = 0.1$ and $\sigma = 0.2$. [Note. Many students find making such conjectures a bit ridiculous: "If a researcher knew such things, why run the study?" Try to convince them that some statistical planning is usually much better than none at all.] In addition, practical matters dictate that at most only 400-425 men can be studied. It is agreed that for the DRUG factor, half of the men will get the placebo, one quarter will get drug A, and one quarter will get drug B. Half of the men will be "required" by their physician to participate in the diet and exercise program.

Last, the tests that Dr. Terr and the statistician agree to focus on are both main effects, the interaction, and three contrasts. One contrast compares the controls against the average of the two drug groups, but only for the men in the diet and exercise program.

3.2 Computing and Tabling the Power Probabilities

This step involves 1) computing a small set of SSH(population) values using one's regular linear models software and 2) using those values to produce power analysis tables.

Consider the SPSS statements below, where the cell codes are DIEX: 1 = "not required," 2 = "required;" DRUG: 1 = placebo, 2 = drug A, 3 = drug B. In ANOVA applications, one easily produces "data" defined by $\mathbf{y}^* = \mathbf{X}_T\beta_T$ by making entering values equal to the scenarios' population cell means: the variables SCNARIO1 and SCNARIO2. BASEN is the base sample size for each cell. The `WEIGHT BY BASEN` statement causes each case to be treated as BASEN cases. In SAS, one uses the FREQ command.

```
TITLE OBTAIN SSH(POP) VALUES FOR HYPOTHETICAL LDL STUDY
DATA LIST LIST/ DIEX DRUG SCNARIO1 SCNARIO2 BASEN
BEGIN DATA
                1    1    -.05     -.05     2
                1    2    -.10     -.12     1
                1    3    -.13     -.18     1
                2    1    -.10     -.12     2
                2    2    -.12     -.15     1
                2    3    -.16     -.20     1
END DATA
WEIGHT BY BASEN
MANOVA SCNARIO1 SCNARIO2 BY DIEX(1,2) DRUG(1,3)/
  CONTRAST(DRUG)= SPECIAL (1 1 1, 2 -1 -1, 0 1 -1)/
        PARTITION(DRUG)=(1 1)/
  DESIGN = DIEX, DRUG, DIEX BY DRUG/
  DESIGN = DIEX = 0, DRUG(1), DRUG(2), DIEX BY DRUG = 0/
```

```
DESIGN = DIEX = 0, DRUG WITHIN DIEX(1) = 0, DRUG(1) WITHIN
         DIEX(2), DRUG(2) WITHIN DIEX(2) = 0/
```

Note that getting the base SSH(population) values for the two scenarios requires the same statements that one would use for regular data analysis. Because this is an unbalanced design, attention must be paid to what fuller and reduced models are used in defining SSH. Here I chose to always let $\mathbf{X}_F$ be the saturated model (rank of $\mathbf{X}_F$ equals number of nonempty cells of design), which is called the "UNIQUE" SSH by SPSS^X MANOVA and the "Type III" SSH in SAS PROC GLM. Other SSH types can be employed, depending on the nature of the design, the research questions, and the philosophy of the data analyst.

The SSH(sample) values obtained from the above input statements are the SSH(population) values based on N = 8. For example, the SSH(population) values for `DRUG(1) WITHIN DIEX(2)` [better labeled "DRUG(2 -1 -1) WITHIN DIEX(2)"] turn out to be .00160 and .00303 for the two scenarios.

Transforming the base SSH(population) values into tables of power probabilities involves the use of a free program, FPOWTAB (F POWer TABles), which has versions written for both the base SAS System (Vers. 5) or for an ANSI-standard FORTRAN 77 compiler. FPOWTAB produces tables showing how the power varies as a function of the "power factors:" 1) scenario for the means, 2) Type I error rate, 3) standard deviation, 4) total sample size, 5) hypothesis tested, and 6) for DFH = 1, two-tailed versus one-tailed alternatives. Total sample size is manipulated by first specifying the basis sample size used to compute the SSH(population) values and then giving a set of multipicative factors ["m" in Equation (2.3)]. Annotated input is given below. [To get these programs send me a PC disk, Macintosh disk, or a BITNET message (PA87458 at node UTKVM1).]

```
'EFFECT OF DIET/EXERCISE AND DRUG THERAPY ON LDL'
8  6
2 'LOW EFFECT'   'BIG EFFECT'
2 .01  .05
2 .1  .2
3 13  26  52
'DIET/EXERCISE MAIN EFFECT' 1 .002   .00288
'DRUG MAIN EFFECT' 2  .00674  .0150 4
'DI/EX BY DRUG INTERACTION'  2 .00034  .00104
'DRUG(2 -1 -1)'  1 .00551 .01201
'DRUG(0 1 -1)'  1 .00123 .00303
'DRUG(2 -1 -1) WITHIN DIEX(2)'  1 .0016 .00303
```

main title (up to 78 chars)
base N, number of cells
scenarios (up to 5)
alpha rates (up to 3)
std devs (up to 3)
base N multipliers (up to 5)

Effects records:
title, DFH, SSH(pop) values

3.3 Evaluating the Power Results

Some FPOWTAB output is given in Table 1. Students find such results easy to understand. Studying the pattern of results also provides concrete ("Numbers!") examples on how changes in the power factors affect power. Some rewarding in-class discussions have been generated by asking students to select the "best" sample size for the study. I impress upon them that this should be Dr. Terr's (now informed) decision.

Table 1: Some Output from FPOWTAB (FORTRAN 77 Version)

```
EFFECT: DRUG(2 -1 -1) WITHIN DIEX(2)
DEGREES OF FREEDOM HYPOTHESIS:    1

SCENARIO: BIG EFFECT
POWERS COMPUTED FROM SSH(POPULATION):        0.0030300
USING THE BASIS TOTAL SAMPLE SIZE:      8
AND TOTAL CELLS IN DESIGN:     6
-----------------------------------------------------------
    STD DEV:      0.1    0.1    0.1    0.2    0.2    0.2
    TOTAL N:      104    208    416    104    208    416
-----------------------------------------------------------
  REGULAR F
ALPHA: .01        .27    .58    .92    .05    .12    .27
ALPHA: .05        .50    .80    .98    .17    .29    .51
  ONE-TAILED T
ALPHA: .01        .36    .68    .95    .09    .18    .36
ALPHA: .05        .63    .88    .99    .25    .40    .63
===========================================================
```

4. The Monte Carlo Game

Within either SAS or SPSS^X one can easily generate normally distributed data with a given mean and variance. For the final part of their individual power analysis projects, students generate and test 10 independent data sets that conform to a situation and hypothesis that has power near .8. Even though they say they "know better," they express surprise when such a powerful case sometimes yields nonsignificant results. I get many "that was fun" comments about this part of the assignment.

5. Log-Linear Models

Given a three-way (I x J x K) table, let π_{ijk} be the population probability in cell ijk, and $\pi = [\, \pi_{111}\ \pi_{112} \ldots \pi_{ijk} \ldots \,]'$ be the vector of the M $(M \leq I \cdot J \cdot K)$ probabilities

satisfying $\pi_{ijk} > 0$. The M-element vector of observed frequencies will then be $\mathbf{y} = [\, y_{111}\ y_{112} \cdots y_{ijk} \cdots]'$. $\mathbf{p} = (1/N)\mathbf{y}$ is the sample analog of π. The true log-linear model for π is

$$\ln(\pi) = \mathbf{X}_T\beta_T\,; \qquad \pi = \exp(\mathbf{X}_T\beta_T)\,.$$

where $\ln(\pi) = [\, \ln(\pi_{111})\ \ln(\pi_{112}) \ldots \ln(\pi_{ijk}) \ldots]'$, $\mathbf{X}_T$ is an M by r_T design matrix having rows $\mathbf{x}'_j$, and $\exp(\mathbf{X}_T\beta_T) = [\, \exp(\mathbf{x}'_1\beta_T)\ \exp(\mathbf{x}'_2\beta_T) \ldots \exp(\mathbf{x}'_M\beta_T)]'$. All design matrices employed are full column rank and have a vector of ones for their first column. Here again we are interested in comparing the fits of a reduced (null) model, $\mathbf{X}_R$, and a "fuller" (alternative) model, $\mathbf{X}_F$: $\mathbf{X}_T \supseteq \mathbf{X}_F \supset \mathbf{X}_R$ with ranks $r_T \geq r_F > r_R$. Let $\mathbf{b}_R$ and $\mathbf{b}_F$ be the maximum likelihood (ML) estimates based on $\mathbf{X}_R$ and $\mathbf{X}_F$, and $\mathbf{y}$. The likelihood ratio (LR) statistic commonly used to compare $\mathbf{X}_R$ and $\mathbf{X}_F$ is

$$G^2(\mathbf{y}, \mathbf{X}_F, \mathbf{X}_R) = 2(\mathbf{y}'\mathbf{X}_F\mathbf{b}_F - \mathbf{y}'\mathbf{X}_R\mathbf{b}_R) = 2N(\mathbf{p}'\mathbf{X}_F\mathbf{b}_F - \mathbf{p}'\mathbf{X}_R\mathbf{b}_R)\,, \qquad (5.1)$$

which I prefer to simply call G^2(sample). Note the similarity of (5.1) and (2.1).

Aickin (1983), relying on the so-called Pitman class of true alternatives, establishes that G^2(sample) is asymptotically distributed as a chi-square random variable with $r_F - r_R$ degrees of freedom and noncentrality parameter,

$$\lambda = G^2(N\pi, \mathbf{X}_F, \mathbf{X}_R) = G^2(\text{population}). \qquad (5.2)$$

Thus any ML/LR log-linear models routine will compute λ if $N\pi$ is supplied as the data. CPOWTAB, the chi-square based analog of FPOWTAB, will compute and table the powers. Monte Carlo studies unequivocally support the approximation (O'Brien, 1986b).

References

Aickin, M. (1983), *Linear Statistical Analysis of Discrete Data*, New York: Wiley.

Cohen, J. (1977), *Statistical Power Analysis for the Behavioral Sciences* (Rev. Ed.), New York: Academic Press.

O'Brien, R.G. (1986a), "Power Analysis for Linear Models," *Proceedings of the Eleventh Annual SAS Users Groups International Conference*, Cary, NC: SAS Institute, Inc., 915-922.

O'Brien, R.G. (1986b), "Using the SAS System to Perform Power Analyses for Log-Linear Models," *Proceedings of the Eleventh Annual SAS Users Groups International Conference*, Cary, NC: SAS Institute, Inc., 778-784.

Address: Statistics Dept., Univ. of TN., Knoxville, TN 37996-0532. BITNET: PA87458 at node UTKVM1.
Support: Faculty Research Fellowship, UTK College of Business Administration.