Power and Sample size calculations for Analysis of Variance

Statistical Consulting Center Division of Statistics University of Idaho

1 Introduction

Power and sample size analyses are important tools for assessing the ability of a statistical test to detect when a null hypothesis is false, and for deciding what sample size is required for having a reasonable chance to reject a false null hypothesis.

Recall that for a test of a statistical hypothesis, the Type I error (α) is the probability of rejecting the null hypothesis when it is true. The Type II error (β) is the probability of not rejecting the null hypothesis when it is false. The power of the test equals 1 - β, and is the probability of rejecting the null hypothesis when it is false. The power will depend on the alternative hypothesis, and we would like to have high power to detect alternative hypotheses of interest.

As an example, we will use a problem that appears in Chapter 5 of the Kirk (1995) Design of Experiments text. It is a study of the effect of sleep deprivation on hand steadiness. In this experiment subjects were randomized to groups receiving either 12, 18, 24, or 30 hours of sleep deprivation. After experiencing the indicated amount of sleep deprivation, each subject filled out a standard machine-generated answer form with small circles that must be filled in, for a given period of time. The response for each subject was the number of circles with pencil marks that strayed outside the boundaries of the circle. The experiment is designed as a completely randomized experiment with 8 subjects in each of the 4 sleep deprivation groups. From past experiments an estimate of the variance in this measurement has been obtained, $\widehat{\sigma^2} = 2.2$.

2 Completely randomized design

2.1 Theory

The completely randomized (CR) design can be expressed as:

$Y_{ij} = \mu + \alpha_j + \varepsilon_{i(j)}$ where $j = 1, ..., p; i = 1, ..., n; \alpha_j = \mu_j - \mu$

and the null hypothesis is: $H_0 : \alpha_j = 0$ for all j. It is also assumed that the errors are identically and independently normally distributed with mean 0 and variance σ^2.

For this example we assume that the treatment effect is fixed. When the null hypothesis is true, the F statistic: F = MSBG/MSWG has an F distribution with p-1 and p(n-1) degrees of freedom. When the null hypothesis is false, the F statistic follows a non-central F distribution with p-1 and p(n-1) degrees of freedom, and noncentrality parameter:

$\lambda = \frac{n \sum \alpha_j^2}{\sigma^2}$

The power of the F test is a monotonically increasing function of the parameter λ. Notice that when the null hypothesis is true, $\lambda = 0$, so that the usual (or central) F distribution is just a special case of the non-central F distribution.

2.2 Calculating noncentrality parameter values

To calculate the power of the ANOVA F test to detect a particular alternative hypothesis, we first compute the critical value for the F statistic using the chosen significance level, α, and the appropriate numerator and denominator degrees of freedom from our sample. We then need to specify the alternative hypothesis in the form of its λ value.

The power of the test to detect the given alternative hypothesis is then equal to the area under the noncentral F distribution to the right of the critical value for the test. Using our notation for the CR example, the critical value for the test is: $\mathsf{F}_{1-\alpha,v_n,v_d}$, where $v_n = p-1$ is the numerator degrees of freedom and $v_d = p(n-1)$ is the denominator degrees of freedom for the F test. Then the power of the test is: $P(F_{v_n,v_d,\lambda} > \mathsf{F}_{1-\alpha,v_n,v_d})$ the probability that the observed statistic having a noncentral F distribution will exceed the critical value of the F test.

2.2.1 Example: The hand steadiness experiment

Scenario 1 (A weak effect)

Suppose that the true mean number of stray circles is 4, 4, 5, and 5, respectively, for the 12, 18, 24 and 30 hour groups. Then

$\mu = (\sum \mu_j)/4 = 18/4 = 4.5$,so then $\alpha_1 = \mu_1 - \mu = 4 - 4.5 = -.5$, and $\alpha_2 = -.5, \alpha_3 = .5$, and $\alpha_4 = .5$, so that $\sum \alpha_j^2 = 1$ and $\lambda = 8(1)/2.2 = 3.6$.

Scenario 2 (A stronger effect)

Suppose that the true mean number of stray circles is 3, 3.5, 4.25, and 6.25, respectively, for the 12, 18, 24 and 30 hour groups. Then the same calculations as above yields $\lambda = 22.3$.

Traditionally tables of the noncentral F distribution are used to calculate the power of the F test, using either λ or the related quantity $\phi = \sqrt{\frac{\lambda}{p}}$. However, now that statistical packages include functions for the noncentral F distribution, there are many advantages to using the computer to calculate power. One reason is that you avoid eye strain reading tiny curves, and another is that you can conveniently produce entire power curves for a whole series of values for λ. For our hand steadiness example, the SAS commands below show the use of the `finv` and `probf` functions to calculate the power of the F test to detect the two given alternative hypotheses:

```
data powerex1 ;
 fcr1 = finv(.95,3,28) ;
 power1 = 1 - probf(fcr1,3,28,3.6) ;
 power2 = 1 - probf(fcr1,3,28,22.3) ;
 run ;
 proc print ; run ;
```

The results are:

```
      The SAS System
OBS FCR1   POWER1  POWER2
 1 2.94669 0.28630 0.97053
```

Thus we have unacceptably low power to detect the weak effect, but very high power to detect the stronger effect of sleep deprivation on hand steadiness.

One obvious question is how to arrive at appropriate values of the noncentrality parameter to use in the power calculations. In the example above, we used an estimate of the variance σ^2 along with some conjectures about the group means μ_j to calculate values of λ . Another approach often mentioned in textbooks is to try to specify a standardized treatment effect:

$d = \frac{\mu_{\max} - \mu_{\min}}{\sigma}$.

There are many possible values of $\sum \alpha_j^2$ for a given $\mu_{\max} - \mu_{\min}$ value, but it can be shown that for a given d, the minimum value of $\sum \alpha_j^2$ is $d^2\sigma^2/2$.

This minimum value yields a λ value of $nd^2/2$, which can be used to calculate conservative power estimates. For our example, if we specified that $d = 1.5$, which roughly corresponds to our stronger effect example, then the value of the noncentrality parameter is $\lambda = 8(1.5)^2/2 = 9$.

Perhaps the best approach to calculating power is to use the ability of the computer to generate an entire power curve (as a function of λ), then calculate some particular λ values and get a feel for the power as a function of λ. For our example, the following SAS code, using a do loop to perform repeated power calculations, will yield a power curve:

```
data powerex2 ;
 do lambda = 0 to 30 by .5 ;
 fcr = finv(.95,3,28) ;
 power = 1 -probf(fcr,3,28,lambda) ;
 output ;
 end ;
 run ;
 proc print ; run ;
 proc plot ; plot power*lambda ; run ;
```

By redrawing the power curve in Proc Insight, you can get a smoother picture than the one drawn by proc plot:

2.3 Sample size calculations

To perform sample size calculations, you will want to specify the part of the noncentrality parameter that does not involve the sample size (n). After specifying $\lambda/n = \sum \alpha_j^2/\sigma^2$, programs can be written that use a do loop where the sample size n is increased every time in the loop. When n increases, then λ increases, and for the CR design also the denominator degrees of freedom $v_d = p(n-1)$ increases. If a specified power is to be achieved, then a condition can be inserted into the do loop to terminate the loop if the calculated power is greater than the specified amount. The following SAS code computes a sample size estimate for a future sleep deprivation experiment, in which the ratio $\lambda/n = \sum \alpha_j^2/\sigma^2$ is set at 1 (a value between the weak and strong effects above), and for which a power of .8 is desired:

```
data powerex3 ;
 do n = 2 to 10000 ;
 ddf = 4*(n-1) ;
 fcr = finv(.95,3,ddf) ;
 lambda = n*(1) ;
 power = 1 -probf(fcr,3,ddf,lambda) ;
 output ;
 if power ge .80 then leave ;
 end ;
run ;
proc print ; title 'Sample size results' ; run ;
```

```
proc plot ; plot power*n ;
 title 'Sample size results' ; run ;
```

Again a nicer plot than that created by proc plot can be created using Proc Insight. Here we have only listed the proc print part of the output:

Sample size results

OBS	N	DDF	FCR	LAMBDA	POWER
1	2	4	6.59138	2	0.10740
2	3	8	4.06618	3	0.18494
3	4	12	3.49029	4	0.26856
4	5	16	3.23887	5	0.35356
5	6	20	3.09839	6	0.43648
6	7	24	3.00879	7	0.51485
7	8	28	2.94669	8	0.58704
8	9	32	2.90112	9	0.65210
9	10	36	2.86627	10	0.70969
10	11	40	2.83875	11	0.75986
11	12	44	2.81647	12	0.80295

These results indicate that with 12 subjects in each of the four groups, we would have power >.80 to detect the specified value of $\sum \alpha_j^2/\sigma^2$.

3 Other designs

Generalizing the power and sample size calculations from the completely randomized design to other experimental designs is very straightforward, but requires the ability to calculate expected values of the mean squares (EMS's) for the model of interest, and also requires the specification of which effects in a model are fixed effects and which are random effects. Although SAS has the capability to calculate the EMS's for experimental designs, for this presentation we will assume that the EMS's are available either from tables in a book, or by hand calculation (Kirk, 1995, Chapter 9).

To calculate power for a test of the form: $F = \frac{MS_n}{MS_d}$, there are 2 cases:

3.1 Fixed effects models

If the effect being tested is from a fixed effects model, then the F statistic again follows a noncentral F distribution with noncentrality parameter:

$\lambda = v_n \left[\frac{E(MS_n)}{E(MS_d)} - 1 \right]$.

As an example, for the completely randomized design,

$E(MS_n) = \sigma^2 + \frac{n \sum \alpha_j^2}{p-1}$, and $E(MS_d) = \sigma^2$, so $\lambda = \frac{n \sum \alpha_j^2}{\sigma^2}$ as stated previously (recall that $v_n = p - 1$ for the CR design).

3.2 Random effects models

If the effect being tested is from a random effects model, then it may not be possible to find a test of the form $F = \frac{MS_n}{MS_d}$, and a quasi-F test may be required. However, if an effect can be tested by an F ratio of the form above, then when the null hypothesis is false the F statistic is distributed as a multiple of a central F

distribution, $\gamma F_{v_n,v_d}$ where $\gamma = E(MS_n)/E(MS_d)$. Thus to calculate power, we use a central F distribution and calculate

$P(F_{v_n,v_d} > \frac{1}{\gamma}\ \mathsf{F}_{1-\alpha,v_n,v_d})$.

In the example below, we are using a completely randomized (two factor) factorial model in which both effects are random. For this model, the expected mean squares of the terms are:

$E(MSA)$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta} + nq\sigma^2_\alpha$
$E(MSB)$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta} + np\sigma^2_\beta$
$E(MSAB)$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta}$
$E(MSWCELL)$	σ^2_ε

In the table above, q is the number of levels of factor B. For the test for the AB interaction, the multiplier $\gamma = E(MSAB)/E(MSWCELL) = 1 + n\sigma^2_{\alpha\beta}/\sigma^2_\varepsilon$. If we can specify a value for the ratio $\sigma^2_{\alpha\beta}/\sigma^2_\varepsilon$ then we can calculate power or sample size. If $p = 3$ and $q = 4$ and we specified that $\sigma^2_{\alpha\beta}/\sigma^2_\varepsilon = .5$, for example, the SAS program below calculates the sample size neccesary to have a power of .8 for the test for AB interaction:

```
data powerex4 ;
 do n = 2 to 10000 ;
 ddf = 12*(n-1) ;
 fcr = finv(.95,6,ddf) ;
 gamma = 1 + n*(.5) ;
 power = 1 -probf(fcr/gamma,6,ddf) ;
 output ;
 if power ge .80 then leave ;
 end ;
 run ;
proc print ; title 'Sample size results' ; run ;
proc plot ; plot power*n ;
 title 'Sample size results' ; run ;
```

Notice that for some tests with random effects (or for fixed effects in a mixed model), that the power of the test can only be made large by increasing the number of levels of a random effect, and not by increasing n. For example, in the model above, the test for the A effect uses F = MSA/MSAB. In that case

$\gamma = E(MSA)/E(MSAB) = 1 + nq\sigma^2_\alpha/(\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta})$,

which only attains large values for large values of q, the number of levels of the B factor.

3.3 Mixed effect models

For mixed effect models, to test either a fixed or a random effect, an F test of the form $F = \frac{MS_n}{MS_d}$ may not be available, thus requiring use of a quasi-F test. This may be a good time to consult a statistician! If , however, a test of the above form is available, then the power of the test can be calculated according to either subsection 1 or 2 above, depending upon whether the effect being tested is a fixed effect or a random effect.

4 References

Kirk, R.E. 1995. Experimental Design: Procedures for the Behavioral Sciences, 3rd edition. Pacific Grove, CA: Brooks/Cole Publishing.