Lecture 20: Power & Sample Size Calculation

Reading Assignment:

- Muller and Fetterman, Chapter 17: "Understanding and Computing Power for the GLM" (for more background)

Motivation

One of the most common questions asked of a statistician about study design is the number of patients to include.

It is an important question, because if a study is too small it will not be able to answer the question posed, and would be a waste of time and money. It could also be deemed unethical because patients may be put at risk with no apparent benefit.

However, it is also undesirable for studies to be too large because resources would be wasted if fewer patients would have sufficed, and it is unethical to expose more patients to the less effective treatment.

Next we will discuss the power calculation for continuous response variables followed by categorical response variables.

Continuous Response Variable

Model Statement

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}$$

with $n >> p$. Assume $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$ and that $\boldsymbol{\beta}$ contains fixed and unknown constants. Also assume that $\mathbf{X}$ contains fixed values, known without appreciable error, conditional upon having collected the sample. Any power computed with the methods described in this chapter applies only to the particular choice of $\mathbf{X}$ used in the calculations. Changing $\mathbf{X}$ changes the design.

Estimation

Assume $\mathbf{X}$ is of full rank. Let

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Then the predicted values can be computed as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},$$

the sum of squares error equals

$$SSE = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}'[I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y},$$

and the variance estimate,

$$\hat{\sigma}^2 = \frac{SSE}{n-p}.$$

Let $\boldsymbol{\beta}$ be the primary parameters and $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$ be the secondary parameters. Assume $\mathbf{C}$ an $a \times p$ matrix ($a \leq p$) with full (row) rank of a. Therefore

$$\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$$

is full (row) rank of a. Thus $\boldsymbol{\theta}$ is both estimable and testable.

The General Linear Hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

$$\text{versus } H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Estimability of $\boldsymbol{\theta}$ and full rank of $\mathbf{M}$ ensures uniqueness of the sum of squares for the hypothesis,

$$SSH = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

and also ensures uniqueness and existence of the likelihood ratio test.

Test Statistic, Null Case

Under the assumptions described above, the maximum likelihood approach provides a test with type I error rate exactly equal to the

target, α. The likelihood ratio test statistic has many equivalent forms:

$$\begin{aligned} F_{obs} &= \frac{SSH/a}{SSE/(n-p)} \\ &= \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\hat{\sigma}^2} \\ &= \frac{\hat{\rho}_*^2/a}{(1-\hat{\rho}_*^2)/(n-p)}, \end{aligned}$$

where $\hat{\rho}_*^2$ represents a generalized squared correlation. When the models spans an intercept and the hypothesis does not span the intercept, then the generalized correlation reduces to the usual multiple correlation ("corrected" for the intercept). Under H_0, $F \sim F(a, n-p)$.

Test Statistic, Non-Null Case

Define $power$ as the probability of rejecting H_0, whether or not H_0 is

true or not. This definition differs slightly from the traditional definition. But this new definition greatly simplifies the discussion of power. With the current definition, if H_0 holds, then the power of the GLH equals the type I error rate. A test is *unbiased* if the expected rejection rate is no more than α for null cases and no less than α for alternative cases. Among all tests unbiased and invariant to location and scale, the likelihood ratio test represents the uniformly most powerful test.

Define

$$\begin{aligned} f_A &= \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)/a}{\sigma^2} \\ &= \frac{\rho_*^2/a}{(1-\rho_*^2)/n} \\ &\approx \frac{\rho_*^2/a}{(1-\rho_*^2)/(n-p)}. \end{aligned}$$

The last equation has no "hats" compared to the last equation in F_{obs}, which uses estimates of parameters. Note that f_A is a parameter, a constant, while F_{obs} is a random variable.

Under H_A,

$$F_{obs} \sim F(a, n-p, \omega),$$

with

$$\omega = a \times f_A = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{\sigma^2}$$

$$= \frac{\rho_*^2}{(1 - \rho_*^2)/n} \qquad (*)$$

Refer ω as the noncentral parameter because it captures the amount by which the model deviates from the central case.

Properties of ω in the GLM

First note that $0 \leq \omega < \infty$. Having $\omega = 0$ implies H_0 holds and power equals α. Also changing the scale of the data (such as from meters to centimeters) does not change ω.

For the independent groups T test of equality of means, assuming

equal cell sizes,

$$\omega = \frac{n}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma^2}.$$

For a multiple regression model that includes an intercept, ω associated with the overall test of (corrected) regression, the usual test of all slopes equal to zero is

$$\omega = \frac{\sigma_Y^2 \rho^2}{\sigma_Y^2(1-\rho^2)/n} = \frac{\rho_*^2}{(1-\rho_*^2)/n}$$

$$= \frac{n\rho_*^2}{(1-\rho_*^2)}$$

Here $\sigma^2 = \sigma_Y^2(1-\rho^2)$ represents the usual residual variance for the model; while σ_Y^2 represents the variance of $\mathbf{Y}$. In turn, ρ^2, the usual squared multiple correlation coefficient, provides a scale free measure of effect and $(1-\rho^2)$ provides a scale free measure of residual variance.

In both special cases, ω increases with sample size, decreases with error variance, and increases with amount of effect. Equation (*) allows generalizing this statement to all cases of GLM. In the general case, $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ captures the size of the effect, while sample size n, hides in $\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ through $\mathbf{X}$. Consider a balanced ANOVA design with cell mean coding, what is $(\mathbf{X}'\mathbf{X})^{-1}$? In summary, ω varies only as function of sample size, mean difference, and error variance.

Computing Power

1. Specify $\alpha, \sigma^2, \mathbf{X}, \boldsymbol{\beta}, \mathbf{C}$ and $\boldsymbol{\theta}_0$.
2. Find the critical value, $f_{crit} = F_F^{-1}(1 - \alpha; a, n - p)$;.
3. Compute the noncentral parameter, ω, as defined in equation (*).
4. Compute $Power = 1 - F_F(f_{crit}; a, n - p, \omega)$.

Factors in Choosing a Design

1. Test size α, which changes when using a Bonferroni correction for multiple analyses
2. The size of σ^2, which must often be varied due to uncertainty about the variable, the population or the study design
3. Varying $\mathbf{X}$ includes changing total sample size, cell size ratios, and the distribution of control variables. Typically balanced designs maximize power.
4. Varying $\boldsymbol{\beta}$ evaluates the impact of the strength and pattern of effects. Some choices of $\boldsymbol{\beta}$ are equivalent in terms of power: for example, for an overall test of equality of means in a three-group ANOVA balanced design, the following choices for $\boldsymbol{\beta}$ are equivalent: $[200, 210, 220]'$, $[0, 10, 20]'$, $[210, 200, 220]'$.
5. The choice of $\mathbf{C}$ plays the primary role in specifying the hypothesis.

6. The choice of $\boldsymbol{\theta}_0$ completes specification of the hypothesis. In most cases, $\boldsymbol{\theta}_0 = \mathbf{0}$.

Using Parameter Estimates in Power Analysis

Unfortunately, several quantities are required before we can do any calculations (and the argument is a little circular!) Speculation drives power analysis not data. Despite that, in some cases data from an earlier study fuel the speculation. Estimation of $\sigma^2, \boldsymbol{\beta}$ or both may be used to compute ω. As a function of one or more parameter estimates, the noncentral value estimate becomes a random variable, as does the corresponding power. The process adds additional source of uncertainty.

Taylor and Muller (1995) proposed the following methods to construct confidence intervals for both noncentrality and power when using $\hat{\sigma}^2$: let

$$\hat{\omega} = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) / a}{\hat{\sigma}^2}$$

with $\hat{\sigma}^2$ based on v degrees of freedom. Also let $c_{crit} = F^{-1}_{\chi^2}(\alpha_c; v)$, the α_c quantile of a χ^2 random variable. Compute the α_c quantile for ω as

$$\hat{\omega}_c = \hat{\omega}\frac{c_{crit}}{v}.$$

Using $\hat{\omega}_c$ to compute power yields an α_c quantile for power. They reported similar results based on estimated σ^2 and $\boldsymbol{\beta}$.

Example1: Kidney Disease

Falk et al (1992) rabdomly assigned 24 participants to one of two treatments intended to slow the worsening of kidney disease. Higher levels of creatinine indicate worse function. Using the reciprocal of serum creatinine level as the dependent variable allowed the investigators to meet the Gaussian assumption. The scientists considered an increase of 0.50dL/mg a clinically important improvement.

If we state the model as

$$\begin{aligned}\mathbf{y}_{24} &= \mathbf{X}_{24\times 2}\boldsymbol{\beta}_{2\times 1} + \mathbf{e} \\ &= \begin{bmatrix} \mathbf{1_{12}} & 0_{12} \\ 0_{12} & \mathbf{1_{12}} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \mathbf{e}\end{aligned}$$

The hypothesis of interest uses $\mathbf{C} = [-1, 1]$ and $\boldsymbol{\theta}_0 = 0$. Let $\delta = \mu_2 - \mu_1$. Then $\boldsymbol{\theta} = \delta = 0.5$, which reflects the investigators' interest. Assuming $\sigma^2 = 0.068$ that estimated from the study, which led to a $\hat{\omega} = 22.06$ and a power of 0.96 under $\alpha = 0.01$. Taylor and Muller computed the 95% CI for ω as [11.01, 36.880] and for power as [0.688, 0.999]. The asymmetry of the χ^2 leads to asymmetric confidence intervals. Taylor and Muller recommended using one-sided confidence intervals for power, since we usually wish to make statements of inequality, such as ensuring power no less than P. The one sided interval is [0.75, 1].

Example2: Medical Cost Suppose we want to design a study to

determine whether there is a linear relationship between nursing home patients' ages and their annual costs for medication. It would be considered unimportant from an economic and medical standpoint if age explained less than .04 of the variability in medication cost. We want a significance level of $\alpha = 0.05$. Calculate the power when sample size n=200.

Solution: $f_{crit} = F_F^{-1}(1-\alpha; a, n-p) = F_F^{-1}(0.95; 1, 198) = 3.889$

$\omega = \frac{n\rho^2}{(1-\rho^2)} = 200 * 0.04/(1-0.04) = 8.33$

$power = 1 - F_F(f_{crit}; 1, n-p, \omega) = 1 - F_F(3.889; 1, 198, 8.33) = 0.82.$

Power Reporting

Single power values rarely suffice to inform the scientist. As statisticians, the authors use tables and plots of power values to return the choices of sample size to the principal investigator. A typical table involves varying two or three of the factors controlling power: mean difference, variance and sample size.

Power for comparing two means

Sample Size per group	Detectable Effect Size	
	80% power	90% power
10	1.3	1.5
20	0.9	1.1
50	0.6	0.7
100	0.4	0.5
200	0.28	0.33

where effect size$= \frac{|\mu_1 - \mu_2|}{\sigma}$.

Dichotomous Responses

For a binary outcome we need to specify type I error, and proportions P1 and P2 where P1 is the expected outcome under the control intervention and P1- P2 is the minimum clinical difference which it is worthwhile detecting.

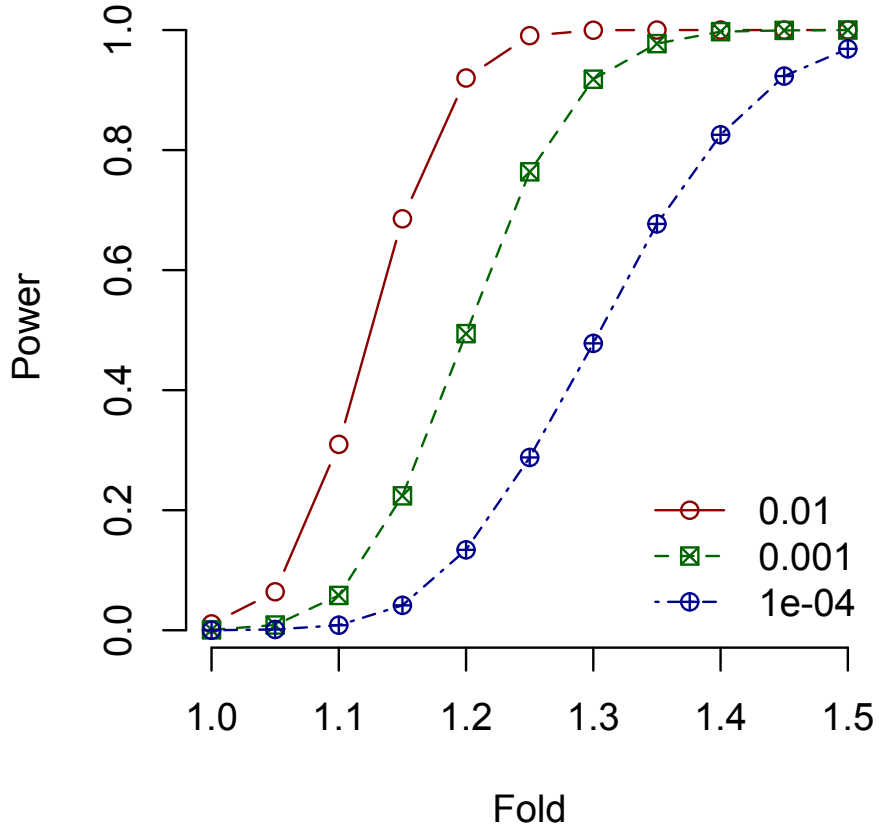
1. Specify $\alpha, P1$ and $P2$.
2. Find the critical value, $z_{crit} = N^{-1}(1 - \alpha/2)$;
3. Calculate $\omega = \frac{P1-P2}{\sqrt{P1(1-P1)/n1+P2(1-P2)/n2}}$
4. Compute $Power = 1 - N(z_{crit}; \omega, 1)$.

Power for comparing two proportions

P1	P2	n1	n2	Power
0.2	0.3	100	100	0.38
0.2	0.3	200	200	0.64
0.15	0.3	100	100	0.73
0.15	0.3	200	200	0.95

Illustrate Power by Figure

Power of differential expression study for three different p-value cutoffs.



Software:

- Commercial one: nquery
  https://uncapps.its.unc.edu/vpn/index.html
- SAS and R have various functions for power analysis
- online power calculation: http://powerandsamplesize.com/

How Much to Do?

In practice the sample size is often fixed by other criteria, such as finance or resources, and the formula is used to determine a realistic effect size. If this is too large, then the study will have to be abandoned or increased in size.

Five key questions regarding sample size:

1. What is the main purpose of the study?

2. What is the principal measure of patient outcome?

3. How will the data be analyzed to detect a treatment difference?

4. What type of results does one anticipate with standard treatment?

5. How small a treatment difference is it important to detect and with what degree of certainty?

Thus in order to calculate the sample size for a study it is first necessary to decide upon what your outcome is. If your outcome variable is continuous you will need to have some measure of what you would expect its mean value to be in the control group together with an estimate of its standard deviation. You will also need to know what size of effect you expect or is desirable (be realistic with this). If your outcome variable is binary you will need to have an idea of the proportions falling into the two outcome categories, and what change in these proportions can be expected or is desirable.

After deciding on the purpose of the study and the principle outcome measure, the investigator must decide how the data are to be summarized and analyzed to detect a treatment difference. Thus, the

investigator must choose an appropriate summary measure of this outcome and then calculate a sample size based on the smallest treatment difference in this summary measure that is of such clinical value that it would be very undesirable to fail to detect. Given answers to all of the five questions above, we can then calculate a sample size.