Selection of Sample Size by Statistical Power[1]

STA441 Spring 2020

[1] See last slide for copyright information.

Background Reading

Optional

- *Data analysis with SAS*, Chapter 8.

How many subjects?

- Mostly, we analyze data sets that people give us. The sample size is given.
- It's better to decide in advance on some rational basis.
- So how many experimental units do you need?
- Need for what?
- Maybe to make a parameter estimate accurate within some designated range.
- Maybe to have a high probability of significant results when a relationship between variables is present.

Testing (null) hypotheses

- Goal is to make correct decisions with high probability.
- When H_0 is true, probability of a correct decision (don't reject) is $1 - \alpha$. That's guaranteed if the model is correct.
- When H_0 if false, we want to reject it with high probability.
- The probability of rejecting the null hypothesis when the null hypothesis is false is called the *power* of the test.
- Power is one minus the probability of a Type II error.
- It is a function of the true parameter values.
- And also the design, including total sample size.

Power is an increasing function of sample size

- Usually, when H_0 is false, larger sample size yields larger power.
- If power goes to one as a limit when H_0 is false (regardless of the exact parameter values) the test is called *consistent.*
- Most commonly used tests are consistent, including the general linear F-test.
- This means that if H_0 is false, you can make the power as high as you wish by making the sample size bigger.

Strategy

- Pick an effect you'd like to be able to detect. An "effect" means a way that H_0 is wrong. It should be just over the boundary of interesting and meaningful.
- Pick a desired power – a probability with which you'd like to be able to detect the effect by rejecting the null hypothesis.
- Start with a fairly small n and calculate the power. Increase the sample size until the desired power is reached.

Distribution theory

- Power depends on the distribution of the test statistic when the null hypothesis is *false*.
- All the distributions you've seen (Z, t, χ^2, F) were derived under the assumption that H_0 is *true*.
- For fixed-effects regression and analysis of variance, the distributions under the *alternative* hypothesis are called *non-central*.
    - The non-central chi-squared.
    - The non-central t.
    - The non-central F.
- These distributions have the same parameters as their ordinary (central) versions, and they also have a *non-centrality parameter*.
- If the non-centrality parameter equals zero, the non-central distribution reduces to the ordinary central distribution.

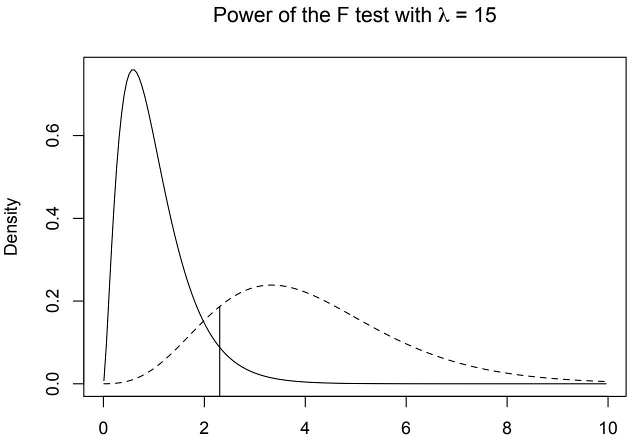
Theorem

If $F^* \sim F(\nu_1, \nu_2, \lambda)$, then

- F^* is stochastically increasing in λ, meaning that for every $x > 0$, $Pr\{F^* > x|\lambda\}$ is an increasing function of λ.
- That is, the bigger the non-centrality parameter, the greater the probability of getting F^* above any point (such as a critical value).
- $\lim_{\lambda \to \infty} Pr\{F^* > x|\lambda\} = 1$.

The greater the non-centrality parameter λ, the greater the power

$\lambda = 0$ means the null hypothesis is true



Non-centrality parameter looks like the test statistic

But with true parameter values instead of estimates

For Testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{h}$, like

$$\left(\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right) \left(\begin{array}{c} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \end{array} \right)$$

We have

$$\begin{aligned} F &= \frac{(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{h})^\prime (\mathbf{C}(\mathbf{X}^\prime \mathbf{X})^{-1}\mathbf{C}^\prime)^{-1} (\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{h})}{r\,MSE} \\ \lambda &= \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})^\prime (\mathbf{C}(\mathbf{X}^\prime \mathbf{X})^{-1}\mathbf{C}^\prime)^{-1} (\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2} \end{aligned}$$

What makes λ big?

$$\lambda = \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})^\prime(\mathbf{C}(\mathbf{X}^\prime\mathbf{X})^{-1}\mathbf{C}^\prime)^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2}$$

- Small σ^2.
- Null hypothesis very wrong.
- Total sample size big.
- Relative sample sizes.
- But sample size is hidden in the $\mathbf{X}^\prime\mathbf{X}$ matrix.

An Important Special Case: Factorial ANOVA with no covariates

- Use cell means coding.
- Skipping a lot of STA305 material …

For cell means coding, λ simplifies to

$$\lambda = n \times (\frac{\mathbf{C}\boldsymbol{\beta} - \mathbf{h}}{\sigma})'(\mathbf{C} \left[\begin{array}{cccc} 1/f_1 & 0 & \cdots & 0 \\ 0 & 1/f_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/f_p \end{array} \right] \mathbf{C}')^{-1}(\frac{\mathbf{C}\boldsymbol{\beta} - \mathbf{h}}{\sigma})$$

- $f_1, \ldots f_p$ are relative sample sizes: $f_j = n_j/n$.
- $\mathbf{C}\boldsymbol{\beta} - \mathbf{h}$ is an *effect*, a particular way in which the null hypothesis is wrong. It is naturally expressed in units of the common within-treatment standard deviation σ, and in general there is no reasonable way to avoid this.
- The magnitude of a difference does not matter, because it depends on the units. What counts is the size of the difference relative to how spread out the data are.

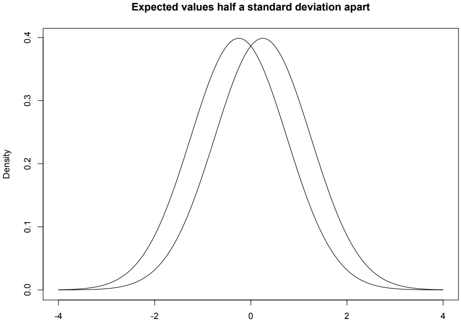
More about the non-centrality parameter

$$\lambda = n \times (\frac{\mathbf{C}\boldsymbol{\beta}-\mathbf{h}}{\sigma})'(\mathbf{C}\left[\begin{array}{cccc} 1/f_1 & 0 & \cdots & 0 \\ 0 & 1/f_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/f_p \end{array} \right]\mathbf{C}')^{-1}(\frac{\mathbf{C}\boldsymbol{\beta}-\mathbf{h}}{\sigma})$$

- Almost always, $\mathbf{h} = \mathbf{0}$.
- The non-centrality parameter is sample size times a quantity that is sometimes called “effect size.”
- The idea is that effect size represents how wrong H_0 is.
- Here, effect size is squared distance between $\mathbf{C}\boldsymbol{\beta}$ and $\mathbf{h}$, in a space that is scaled and stretched by the relative sample sizes – an aspect of design.

Example: Comparing two means

Suppose we have a random sample of size n_1 from a normal distribution with mean μ_1 and variance σ^2, and independently, a second random sample from a normal distribution with mean μ_2 and variance σ^2. We wish to test $H_0 : \mu_1 = \mu_2$ versus the alternative $H_1 : \mu_1 \neq \mu_2$. If the true means are a half a standard deviation apart, we want to be able to detect it with probability 0.80.



Two-sample t-test, or F-test

Non-central t or non-central F

- We'll use F.
- Skipping the derivation, get

$$\lambda = nf(1-f)d^2$$

where $f = \frac{n_1}{n}$ and $d = \frac{|\mu_1 - \mu_2|}{\sigma}$.

$\lambda = nf(1-f)d^2$, where $f = \frac{n_1}{n}$ and $d = \frac{|\mu_1 - \mu_2|}{\sigma}$

- For two-sample problems, d is usually called effect size. The effect size specifies how wrong the null hypothesis is, by expressing the absolute difference between means in units of the common within-cell standard deviation.
- The non-centrality parameter (and hence, power) depends on the three parameters μ_1, μ_2 and σ^2 only through the effect size d.
- Power depends on sample size, effect size and an aspect of design – allocation of relative sample size to treatments.
- Equal sample sizes yield the highest power in the 2-sample case.

Back to the problem

$\lambda = nf(1-f)d^2$

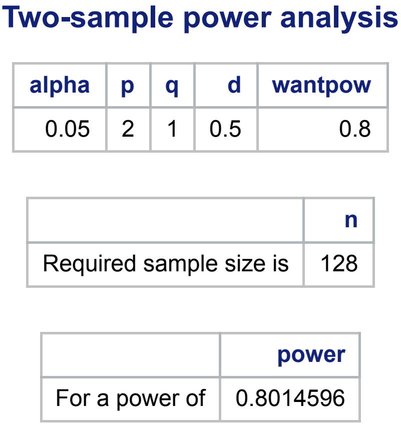
We wish to test $H_0 : \mu_1 = \mu_2$ versus the alternative $H_1 : \mu_1 \neq \mu_2$. If the true means are a half a standard deviation apart, we want to be able to detect it with probability 0.80.

$$\begin{aligned} \lambda &= nf(1-f)\left(\frac{|\mu_1 - \mu_2|}{\sigma}\right)^2 \\ &= n\frac{1}{2}\left(1-\frac{1}{2}\right)\left(\frac{1}{2}\right)^2 \\ &= \frac{n}{16} \end{aligned}$$

SAS proc iml

```
/*****************  fpow1.sas *********************/
title 'Two-sample power analysis';

proc iml; /* Replace alpha, q, p, d and wantpow below    */
     alpha = 0.05;  /* Signif. level for testing H0: C Beta = t  */
     q = 1;         /* Numerator df = # rows in C matrix         */
     p = 2;         /* There are p beta parameters               */
     d = 1/2;       /* d = |mu1-mu2|/sigma */
     wantpow = .80; /* Find n to yield this power                */
     power = 0; n = p; oneminus = 1-alpha; /* Initializing ... */
     do until (power >= wantpow);
        n=n+1 ;
        ncp = n * 1/4 * d**2;
        df2 = n-p;
        power = 1-probf(finv(oneminus,q,df2),q,df2,ncp);
     end;
     print alpha p q d wantpow;
     print "Required sample size is " n;
     print "For a power of " power;
/**********************************************************
```

Output



Two-sample power analysis

alpha	p	q	d	wantpow
0.05	2	1	0.5	0.8

	n
Required sample size is	128

	power
For a power of	0.8014596

To do a power analysis for any factorial design

$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{h}$

All you need is a vector of relative sample sizes and a vector of numbers representing the differences between $\mathbf{C}\boldsymbol{\beta}$ and $\mathbf{h}$ in units of σ.

$$\lambda = n \times \left(\frac{\mathbf{C}\boldsymbol{\beta} - \mathbf{h}}{\sigma} \right)' \left(\mathbf{C} \left[\begin{array}{cccc} 1/f_1 & 0 & \cdots & 0 \\ 0 & 1/f_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/f_p \end{array} \right] \mathbf{C}' \right)^{-1} \left(\frac{\mathbf{C}\boldsymbol{\beta} - \mathbf{h}}{\sigma} \right)$$

Example: Test for interaction

	Level of B		
Level of A	1	2	Average
1	μ_{11}	μ_{12}	$\mu_{1.}$
2	μ_{21}	μ_{22}	$\mu_{2.}$
3	μ_{31}	μ_{32}	$\mu_{3.}$
Average	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

$$H_0 : \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22} = \mu_{31} - \mu_{32}$$

$H_0 : \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22} = \mu_{31} - \mu_{32}$

$$\begin{array}{cccc}
\mathbf{C} & \boldsymbol{\beta} & = & \mathbf{h} \\
\begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix} & \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \end{pmatrix} & = & \begin{pmatrix} 0 \\ 0 \end{pmatrix}
\end{array}$$

Suppose this null hypothesis is false in a particular way that we want to be able to detect.

Null hypothesis is wrong

Suppose that for $A = 1$ and $A = 2$, the population mean of Y is a quarter of a standard deviation higher when $B = 2$, but if $A = 3$, the population mean of Y is a quarter of a standard deviation higher for $B = 1$. Of course there are infinitely many sets of means satisfying these constraints, even if they are expressed in standard deviation units. But they will all have the same effect size. One such pattern is the following.

	Level of B	
Level of A	1	2
1	0.000	0.250
2	0.000	0.250
3	0.000	-0.250

Sample sizes are all equal, and we want to be able to detect an effect of this magnitude with probability at least 0.80.

All we need is true $\mathbf{C}\boldsymbol{\beta}$

$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is wrong.

	Level of B	
Level of A	1	2
1	0.000	0.250
2	0.000	0.250
3	0.000	-0.250

$$\mathbf{C}\boldsymbol{\beta} = \left(\begin{array}{r} 0 \\ -0.5 \end{array} \right)$$

Matrix calculations with proc iml

$\lambda = \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2}$

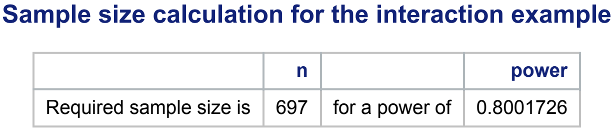
```
/*****************  fpow2.sas *********************/
title ’Sample size calculation for the interaction example’;

proc iml;
/********* Edit this input: Rows of matrices are separated by commas  ********/
     alpha = 0.05; wantpow = .80;
     f = {1,1,1,1,1,1};          /* Relative sample sizes       */
     C = { 1 -1 -1  1  0  0,       /* Contrast matrix             */
           0  0  1 -1 -1  1};
     eff = {0, 0.5};    /* In standard deviation units */
/*****************************************************************************/
```

fpow2.sas continued

$\lambda = \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2}$

```
     p = nrow(f) ; q = nrow(eff); f = f/sum(f);
     core = inv(C*inv(diag(f))*C`);
     effsize = eff`*core*eff;
     power = 0; n = p; oneminus = 1-alpha;  /* Initializing ...*/
     do until (power >= wantpow);
        n = n+1 ;
        ncp = n * effsize;
        df2 = n-p;
        power = 1-probf(finv(oneminus,q,df2),q,df2,ncp);
     end; /*  End Loop */
     print " Required sample size is " n " for a power of " power;
/**********************************************************
```

Output



Sample size calculation for the interaction example

	n		power
Required sample size is	697	for a power of	0.8001726

$697/6 = 116.1667$ and $117 * 6 = 702$, so a total of $n = 702$ experimental units are needed for equal sample sizes and the desired power.

Beyond the F-tests

- Lots of large-sample chi-squared tests (like Wald and Likelihood Ratio) have limiting non-central chi-squared distributions under H_1.
- Non-centrality parameters look like the test statistics.
- For unfamiliar tests, simulation may be easier.

Copyright Information