Power and Sample Size

Tim Hanson

Department of Statistics
University of South Carolina

February, 2017

Modified from originals by Gary W. Oehlert

Background

We have already defined type I and II errors.

	Reality/State of nature	
Decision	Null correct	Null false
Fail to reject	☺	Type II error
Reject	Type I error	☺

The type I error rate $\mathcal{E}$ is easy to set, we just choose it.

Power is the probability of rejecting the null when the null is false. Power is the probability of declaring a difference when the difference is there (getting that lower right smiley).

Power is a much more difficult customer than $\mathcal{E}$.

Well-designed experiments are efficient

You should design your experiments to have “appropriate” power.

- If the power is too low, then you’re just wasting your time and resources running an experiment with no chance of finding what you are looking for.
- If the power is too high, then you are spending resources in this experiment that might be better spent somewhere else.

Appropriate power is probably in the .7 to .95 range, but it is situationally dependent.

Power depends on lots of unknown things

Power for the F test comparing the separate means model with the single mean model depends on practically everything:

- The type I error rate $\mathcal{E}$.
- The numerator and denominator degrees of freedom for the F test; these obviously depend on g and N for the separate means model. We will only consider the separate means model to start with.
- The "non-centrality parameter ζ, which itself depends on the sample sizes $n_1, \ldots, n_g$, the non-null treatment means $\mu_1, \ldots, \mu_g$, and the error variance σ^2.

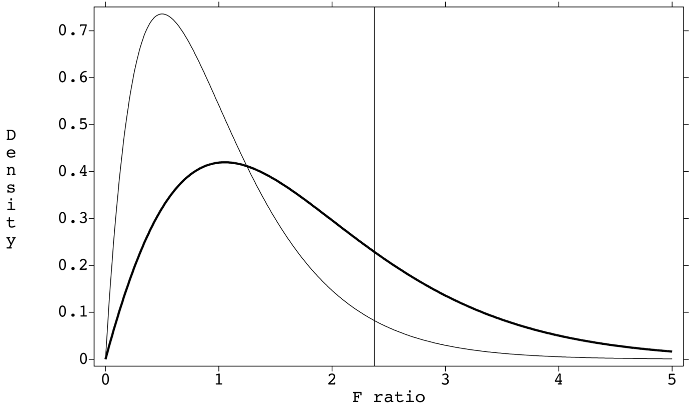
Central and non-central F distributions

Under the null, the F statistic follows a central F dist'n with $g-1$ and $N-g$ df.

Combine this with $\mathcal{E}$ and we get a critical value: reject for F statistics larger than the critical value. (Equivalently, any F in that range will have a p-value less than $\mathcal{E}$.)

When the null is false, the F statistic follows a non-central F distribution with $g-1$ and $N-g$ df. The distribution is shifted to the right, and ζ controls the amount of shift to the right.

Probability of being to the right of the critical value is the power. As you decrease $\mathcal{E}$, it becomes more difficult to reject the null (that moves the critical value to the right so you need a bigger F statistic to reject). For fixed g, N, and ζ, smaller $\mathcal{E}$ leads to lower power.


Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
F ratio

Non-centrality parameter ζ

Let $\mu_i = \mu + \alpha_i$ where we use the α_is with $\sum_i n_i\alpha_i = 0$. Then

$$\zeta = \frac{\sum_{i=1}^{g} n_i\alpha_i^2}{\sigma^2}$$

The expected value of the MS_{Trt} is

$$E[MS_{Trt}] = \sigma^2 + \frac{\sum_{i=1}^{g} n_i\alpha_i^2}{g-1} = \sigma^2\frac{g-1+\zeta}{g-1}$$

Recall $E[MS_E] = \sigma^2$ and $F = \frac{MS_{Trt}}{MS_E}$. ζ is a measure of how far the alternative state of nature is from the null.

- ζ increases if you increase the sample sizes.
- ζ increases if the error variance is smaller.
- ζ increases if the means μ_i are farther apart.

Excuse me, but . . .

The discerning student will remark that ζ depends on lots of stuff we don't know, like the μ_is and σ^2. *If we knew the μ_is, we wouldn't be doing the experiment in the first place!* So what gives?

In practice, power analysis and sample size selection are a big exercise in "Let's pretend" or "What if?"

We can control $\mathcal{E}$, and we can control $n_1, \ldots, n_g$, but otherwise we are plugging in some hypothesized means and error variance and asking what the power would be for that state of nature.

Think about alternatives

To make power analysis useful, you must be able to specify some scientifically or practically meaningful set of alternative means μ_i, and you must be able to make a guess as to how large the error variance σ^2 is.

Find alternative means where you can say, "If this were true, I would want to know about it," and then design for those interesting alternatives.

Examples of alternatives

Examples might be

- A doubling of the mutation rate is practically significant, so I want to design for that.
- An increase in MPG of 1 is relevant, so I will design for that.
- A 20% reduction in the serum concentration of a hormone is diagnostic, so I design for that.

Most granting agencies will require a power analysis before funding a proposal.

Estimating σ^2

OK, but what about σ^2? Some possibilities include:

- Variance from a pilot study.
- Variance from similar experiments in your lab or in the literature (most likely in my experience).
- Theoretical variances (possible for binomial counts and some other situations).
- Analytical variance of equipment (generally an underestimate of σ^2).

It's probably best to do multiple power analyses that cover a range of plausible σ^2 values.

Smallest interesting difference D

Suppose that you have equal sample sizes $n_1 = \cdots = n_g = n$, and you think that any configuration of means where two means are D or more units apart is interesting.

The smallest value of ζ for that description is

$$\zeta_0 = \frac{nD^2}{2\sigma^2}$$

Any ζ for two means D units apart with sample sizes n will be at least as big as ζ_0.

Thus the power for any of the other ζs will be at least as big as what you compute for ζ_0.

Power in R

`power.anova.test` in the `cfcdae` package allows you to enter either

- ζ, the F(df_1, df_2) distribution to use, and Type I error α. Note that $df_1 = g - 1$ and $df_2 = N - g$ for oneway ANOVA.
- Actual means $\mu_1, \ldots, \mu_g$, sample sizes $n_1, \ldots, n_g$, σ^2, and α. This one is easier but the first one can be used in more complex situations.

Say $n_1 = n_2 = n_3 = 5$, $\sigma^2 = 4$ and $\mu_1 = 10$, $\mu_2 = 11$, and $\mu_3 = 15$. Then $\mu^* = 12$ and $\zeta = \frac{1}{\sigma^2}\sum_{i=1}^{g} n_i\alpha_i^2 = \frac{1}{4}[5(10-12)^2 + 5(11-12)^2 + 5(15-12)^2] = \frac{5}{4}[4+1+9] = 17.5$.

```
> power.anova.test(ncp=17.5,df1=2,df2=12,alpha=.05)
[1] 0.9170125
> power.anova.test(means=c(10,11,15),ns=c(5,5,5),sigma2=4,alpha=.05)
[1] 0.9170125
```

The second way is easier for oneway ANOVA; use it for your homework.

Sample size

You have chosen $\mathcal{E}$, you have some interesting values for the μ_is, and you have a pretty good idea what σ^2 is.

Sample size analysis takes those and finds the smallest sample sizes n_i that will achieve a specified level of power.

In principle this involves computing power for a lot of different sample sizes and finding the one that is just big enough. In practice, we just use R.

Sample size in R

`sample.size.anova` in the `cfcdae` package allows you to enter either

- The smallest power needed, Type I error α, the number of groups g, and ζ for $n_1 = \cdots = n_g = 1$, i.e. $\zeta = \frac{1}{\sigma^2}\sum_{i=1}^{g} \alpha_i^2$. Note that $df_1 = g - 1$ and $df_2 = N - g$ for oneway ANOVA.
- The smallest power needed, Type I error α, the actual means $\mu_1, \ldots, \mu_g$, and σ^2.

Sample size in R

Say $\sigma^2 = 4$ and $\mu_1 = 10$, $\mu_2 = 11$, and $\mu_3 = 15$. Then $\mu^* = 12$ and ζ for $n_1 = n_2 = n_3 = 1$ is
$\zeta = \frac{1}{\sigma^2}\sum_{i=1}^{g} n_i\alpha_i^2 = \frac{1}{4}[(10-12)^2 + (11-12)^2 + (15-12)^2] = \frac{1}{4}[4+1+9] = 3.5.$

```
> sample.size.anova(.95,.05,ncp1=3.5,ngrps=3)
$nis
[1] 6 6 6
$power
[1] 0.9665357
> sample.size.anova(.95,.05,means=c(10,11,15),sigma2=4)
$nis
[1] 6 6 6
$power
[1] 0.9665357
```

The second way is easier for your homework problems.

Confidence intervals

Another approach to sample sizes picks n so that confidence intervals are short enough.

For a contrast, we use the CI

$$\sum_{i=1}^{g} w_i \overline{y}_{i\bullet} \pm t_{\mathcal{E}/2,\nu} \sqrt{MS_E \sum_{i=1}^{g} \frac{w_i^2}{n_i}}$$

The margin of error is thus

$$MOE = t_{\mathcal{E}/2,\nu} \sqrt{MS_E \sum_{i=1}^{g} \frac{w_i^2}{n_i}}$$

where ν is the df for MSE. The width of the interval is $W = 2 \times MOE$.

Assuming $n_1 = \cdots = n_g = n$

If we assume that the n_is are all equal, we can solve to get:

$$n \approx \frac{t^2_{\mathcal{E}/2,\nu} MS_E \sum_{i=1}^{g} w_i^2}{MOE^2}$$

We haven't done the experiment yet, so we don't know MS_E, and we will instead use a guess of σ^2 as we did in power analysis.

Confidence intervals

We know our desired MOE, we know the w_is, we have a guess for σ^2 which we use as a guess for MS_E.

Compute n_0 by substituting a normal percent point for the t-percent point.

$$n_0 \approx \frac{(\Phi^{-1}(1-\mathcal{E}/2))^2 \sigma^2 \sum_{i=1}^{g} w_i^2}{MOE^2}$$

This gives you a starting point. Now start n at n_0 and increment it until

$$n \geq \frac{t^2_{\mathcal{E}/2, g(n-1)} \sigma^2 \sum_{i=1}^{g} w_i^2}{MOE^2}$$