Statistical Power

Paul Gribble

Winter, 2019

Statistical Power

- power is the ability of a statistical test to detect real differences when they exist
- β is the probability of failing to reject the null hypothesis when it is in fact false (Type-II error)
- β is the probability of failing to reject the restricted model when the full model is a better description of the data, even with the requirement to estimate more parameters

$$\text{power} = 1 - \beta$$

- power is the probability of rejecting the null hypothesis when it is in fact false

Type-I vs Type-II error & hypothesis testing outcomes

		Reality	
		H_0 is true	H_1 is true
Research	H_0 is true	Accurate $(1-\alpha)$	Type-II error (β)
	H_1 is true	Type-I error (α)	Accurate $(1-\beta)$

Statistical Power

- how sensitive is a given experimental design?
- how likely is our experiment to correctly identify a difference betweeen groups when there actually is one?
- what sample size is required to give an experiment adequate power?
- how many subjects do we need to include in each group sample?

Effect Size

- we need some way of assessing the expected size of the effect we are proposing to detect
- one measure is the standardized measure of effect size, f

$$
\begin{aligned}
f &= \sigma_m / \sigma_\epsilon \\
\sigma_m &= \sqrt{\frac{\sum(\mu_j - \mu)^2}{a}} = \sqrt{\frac{\sum \alpha_j^2}{a}} \\
\mu &= \left(\sum_j \mu_j \right) / a \\
\sigma_\epsilon &= \text{within-group standard deviation}
\end{aligned}
$$

Effect Size

- If you have pilot data you can compute values for f
- If not, Cohen (1977) suggests the following definitions:
  - "small" effect: $f = 0.10$
  - "medium" effect: $f = 0.25$
  - "large" effect: $f = 0.40$
- so for medium effect, standard deviation of population means across groups is $1/4$ of the within-group sd

Power Charts

- Cohen (1977) provides tables that let you read off the power for a particular combination of numerator df, desired Type-I error rate, effect size f, and subjects per group
- four factors are varying — tables require 66 pages!
  - seriously
- It's 2019, Let's use `R` instead
  - `power.t.test()`
  - `power.anova.test()`

An example

- e.g. you are planning a reaction-time study involving three groups ($a = 3$)
- pilot research & data from literature suggest population means might be 400, 450 and 500 ms with a sample within-group standard deviation of 100 ms
- suppose you want a power of 0.80 — how many subjects do you need in each sample group?

An example

```
power.anova.test(groups=3, n=NULL,
  between.var=var(c(400,450,500)),
  within.var=100**2, sig.level=0.05,
  power=0.80)


     Balanced one-way analysis of variance power calculatio

         groups = 3
              n = 20.30205
    between.var = 2500
     within.var = 10000
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

...but since we know how to program in R

- simulate! Simulate sampling from two populations
  - whose means differ by the expected amount
  - whose variances are a particular value
  - postulate a particular sample size N
- sample and do your statistical test many times (e.g. 1000) and see what proportion of times you successfully reject the null (your power)
- If power is not high enough, try a larger sample size N and repeat. Keep increasing N in simulation until you get the power you want
- computationally intensive, but allows you to test any experimental situation that you can simulate

Cautionary note: calculating "observed power" after rejecting the null

- you run an experiment, do stats, and end up failing to reject H_0
- two possibilities:
  1. there is in fact no difference between population means, and your experiment correctly identifies this
  2. there is a difference, but your experiment is not statistically powerful enough to detect it (for e.g. because within-group variability is high)
- can we use power calculations to see if we "had enough power" to detect the difference?
- no — not appropriate use of power analysis (although frequently taught)

Hoenig & Heisey (2001)

- doing a power analysis after an experiment that failed to reject the null, to see if "there was enough power" to detect the difference, is inappropriate
- the result of a post-hoc power analysis is completely redundant with the probability (p-value) obtained in the original analysis
- one can be obtained directly from the other
- you don't learn anything new by doing a post-hoc power analysis
- See Hoenig & Heisey (2001) for the full story

Challenges of power analyses

- you must have estimates of expected difference between means
- you must have estimates of within-group variability
- computing power for more complex experimental designs can be complicated — see Maxwell & Delaney text for examples