Power Analysis

for user experiments

Power Analysis

My goal:

Teach how to scientifically decide whether a sample size is sufficient for a certain study

My approach:

- Quick review of effect sizes, p-values, and power
- Intro to power analysis
- Demo using G*Power

Slides

These slides are based on an excellent guide written by the Institutional Review Board at Utah State University

https://rgs.usu.edu/irb/files/uploads/A_Researchers_Guide_to_Power_Analysis_USU.pdf

A quick review

of effect sizes, p-values, and power

A quick review

Whenever you do stats, you should report (at least) two things*:

- effect size
- p-value

Effect size & p-value

Effect size: the strength of a result

does not depend on sample size!

P-value: the likelihood that the effect was due to chance

very much depends on sample size!

Example

Do married men weigh more than single men?

Find two random married men: both around 180 lbs

Find two random single men: both around 160 lbs

Effect size: 20 lbs

Significant?

No way, sample size way too small!

Example

Let's increase the sample size to 8:

$N_m = 4$, $Mean_m = 182$, $SD_m = 15$

$N_s = 4$, $Mean_s = 170$, $SD_s = 15$

Effect size: 12 lbs

Is this a large effect? —> Need to standardize it!

Cohen's d = ($Mean_m$ – $Mean_s$)/pooled SD

(182–170)/15 = 0.8... this is indeed a large effect

Is it significant? Again, no!

Reflection

Small studies (N << 100) may find medium or large effects that are not significant

Waste of resources! (unless they are pilot studies)

Large studies (N >> 100) may find very small effects that are significant

Also a waste of resources! (could have done with fewer)

How can we prevent wasting resources?

Do a power analysis!

Power analysis

an introduction

Power analysis

A calculation involving the following 4 parameters:

- Alpha (cut-off p-value, often .05)
- Power (probability of finding a true effect, often .80 or .85)
- N (sample size, usually the thing we are trying to calculate)
- Effect size (usually the “expected effect”)

Types

A priori: compute N, given other variables

Conducted before you run your study

Post-hoc: compute power, given other variables

Conducted afterwards to find out if you had enough participants to detect the found effect

Sensitivity: compute effect size, given other variables

Find out the minimum effect size you can detect, given the number of participants

Alpha and power

	There is a real effect	There is no real effect
Found an effect	**Power**	alpha (false positive)
Found no effect	beta (false negative)	1–alpha (true negative)

Expected effect

An “educated guess” based on:

- Pilot study results
- Findings from similar studies
- Whatever is considered “meaningful”
- Educated guess

Examples per test

Statistic	Small	Medium	Large
Means - Cohen's d	0.2	0.5	0.8
ANOVA - Cohen's f	0.1	0.25	0.4
ANOVA - eta squared	0.01	0.06	0.14
Regression - f-squared	0.02	0.15	0.35
Correlation - r or point biserial	0.1	0.3	0.5
Correlation - R-squared	0.01	0.06	0.14
Association - 2x2 odds ratio	1.5	3.5	9
Association - w or Phi	0.1	0.3	0.5

Calculations

What if the effect size is not provided in similar studies?

Compute it!

Comparing means (e.g. t-test): Cohen's d:

$(Mean_a - Mean_b)$/pooled SD

or: $2t/\sqrt{(df)}$

Calculations

Anova: eta-squared and Cohen's f:

eta-squared = $(f)^2$ = SSm/SSt

Note: SPSS reports the *partial* eta-squared!

can also be used, but different (difficult) calculation

Regression: f-squared:

f^2 = partial R^2/(1-partial R^2)

or: calculate from ANOVA table (SSm/SSt)

G*Power demo

power analysis made easy!

G*Power demo

An existing study found that a new TurboTax interface reduced tax filing time from 3.0 hours (SD: 0.5 hours) to 2.7 hours (SD: 0.5 hours).

You created a new interface that you think is even better. How many participants do you need to find an effect that is at least the same size? (assume 85% power)

G*Power demo

You conducted a linear regression testing the effect of number of previous privacy violations on 35 Facebook users' privacy concerns (controlling for age and gender).

The number of previous violations was not significant.

The model without this variable had an R^2 of 0.15.

The model with this variable had an R^2 of 0.30.

What was your power? What sample size should you use to find an effect of this size with 85% power?

G*Power demo

You want to test the combined effect of 6 text sizes and 6 background colors on text readability. You only have money for 150 study participants.

What is the maximum effect size you can find (with 85% power) for a main effects of text size and background color?

What about the interaction effect?

Would it help if you only test 2 sizes and colors?

Final thoughts...

a few warnings, and a final cool trick...

Final thoughts...

Your Mileage May Vary!

Because power cannot be 100%, there is no guarantee you will find an effect!

The effect in your study might be smaller than in previous work!

Your may need to exclude faulty/outlying participants!

Better to estimate conservatively!

Or check out the graphs to see what would happen...

Tiny samples

Be aware of tiny samples (even when they report significant results)

- Randomization doesn't work well in tiny samples
- Tiny samples fall prey to the "publication bias"
- Due to the "winner's curse", tiny samples overestimate the real effect size

These problems are worst for counter-intuitive results

- Ask your friendly neighborhood Bayesian statistician

A final cool trick...

Let's say you need to collect 150 participants...

Ugh... 3 weeks of my time!

...why not run a quick analysis after the first 50 to see if the results are significant?

That's called "p-hacking", and is not allowed

Why? Because you inflate alpha by "peeking"

But what if you compensate by reducing your alpha?

That's allowed! It's called sequential analysis

A final cool trick...

After 50 participants, you do an analysis

3 options:

- No significance, low effect size (reaction: abandon study)
- Significant result (reaction: stop study, take 2 weeks off)
- No significance, but decent effect size (reaction: continue collecting data)

See http://dx.doi.org/10.1002/ejsp.2023 for more details...

“It is the mark of a truly intelligent person
to be moved by statistics.”

T H A N K S !

George Bernard Shaw