OPAM workshop: “Am I underpowered?”

Chris Jungerius

2025 – 07 - 03

Intro

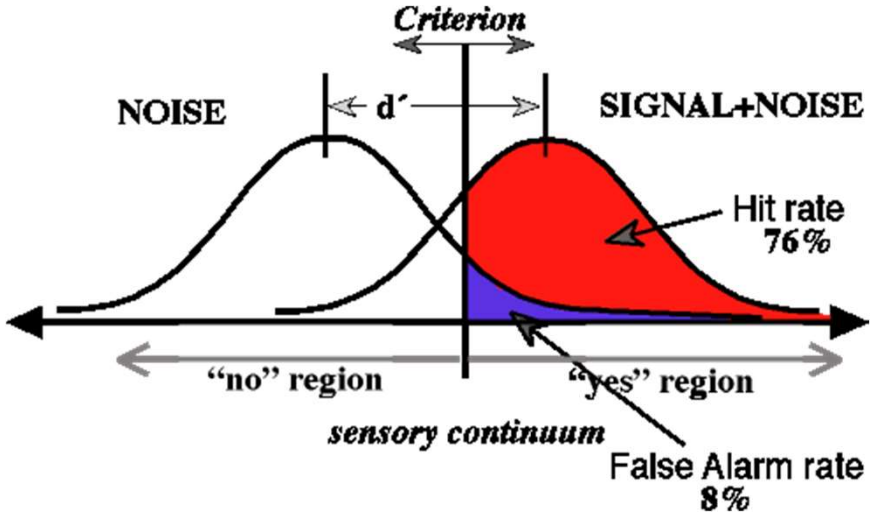
- What is power again?
- Why should I care?
- How do I perform a power analysis?

The NHST framework

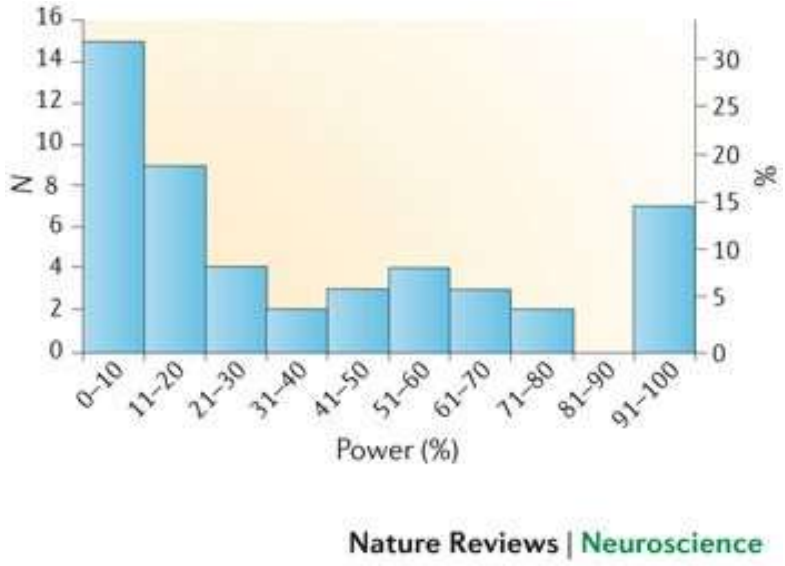
- Falsification in a noisy world: true counterfactuals become impossible
- Need a *decision rule* to reject H0
- NHST: use *sampling distribution* of test statistics under H0 →long-run error control

The NHST framework

- Decision rules for H0 rejection give us long-run error control
- Two types of errors: type I/false positive, and type II/false negative
- Probability of false positive *alpha* (typically 0.05)
- Probability of false negative *beta* (typically 0.2)
- Power = 1 – beta = *ability to correctly reject H0 if there is an effect*



Neuroscience & Psychology have a power problem



Button et al. (2013)

Neuroscience & Psychology have a power problem

		Small effect		Medium effect		Large effect	
Subfields or other surveys	**Records/Articles**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Cognitive neuroscience	7,888/1,192	0.11	0.14	0.40	0.44	0.70	0.67
Psychology	16,887/2,261	0.16	0.23	0.60	0.60	0.81	0.78
Medical	2,066/348	0.15	0.23	0.59	0.57	0.80	0.77
All subfields	26,841/3,801	0.11	0.17	0.44	0.49	0.73	0.71
Cohen (1962)	2,088/70	0.17	0.18	0.46	0.48	0.89	0.83
Sedlmeier & Gigerenzer (1989)	54 articles	0.14	0.21	0.44	0.50	0.90	0.84
Rossi (1990)	6,155/221	0.12	0.17	0.53	0.57	0.89	0.83
Rossi (1990); means of surveys	25 surveys		0.26		0.64		0.85

doi:10.1371/journal.pbio.2000797.t001

Szucs & Ioannidis (2017)

What are the problems with low power studies?

- Obviously, if you are underpowered, the chance that you miss a true effect is larger
- If you do find an effect in an underpowered study, the effect size you'll find is most likely inflated (sometimes severely!)
- If a study is underpowered, the ratio of true positives to false positives it generates goes down: significant findings are less likely to be "real"!

So what determines power?

- Size of effect
- Noise
- Size of sample

- *Power is one of the main reasons why we care about sample sizes!*

So what does this mean for me?

- Running well-powered studies is essential for making your inferences precise and reproducible.
- We must think deeply about our study design (and particulary our sample size) before we run our study!

Side-note: post-hoc (“observed”) power

- Often asked for by reviewers
- Useless for assessing the power of your study: deterministic function of your p-value
- The only way is “prospective”: is/was my design capable of detecting effects of a given size?

Approaches to power analysis

Analytical and Simulation-based

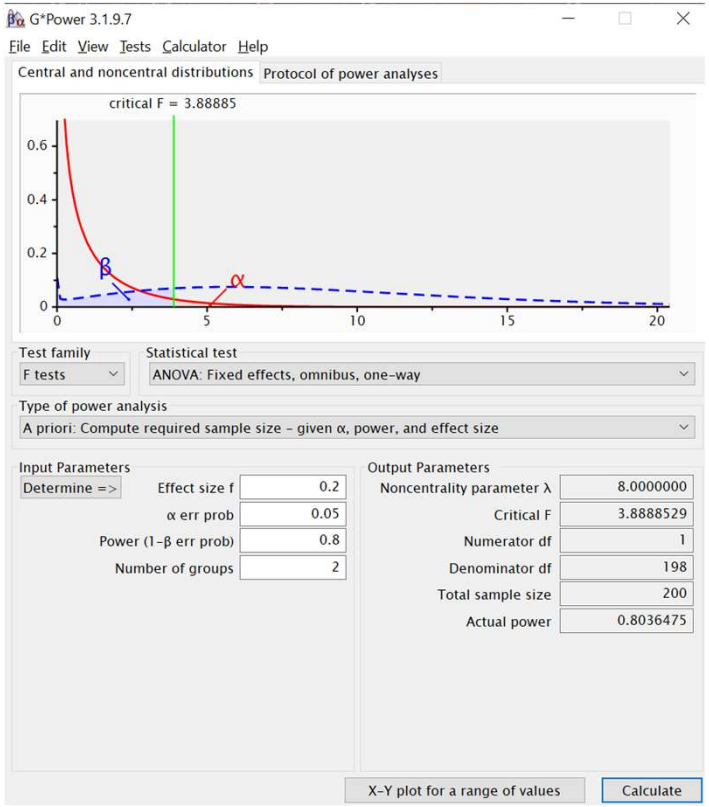
Analytical approach: Just look it up!

Traditional, frequentist analysis ($p < .05$)

	d = .4	d = .5	d = .6
1 variable between-groups			
• 2 levels	**200**	130	90
• 2 levels, null hypothesis	**860**	860	860
• 3 levels (I = II > III)	**435**	285	195
• 3 levels (I > II > III)	**1740**	1125	795
1 variable within-groups			
• 2 levels	**52**	34	24
• 2 levels, null hypothesis	**215**	215	215
• 3 levels (I = II > III)	**75**	50	35

Brysbaert (2018)

Analytical approach: Just look it up!



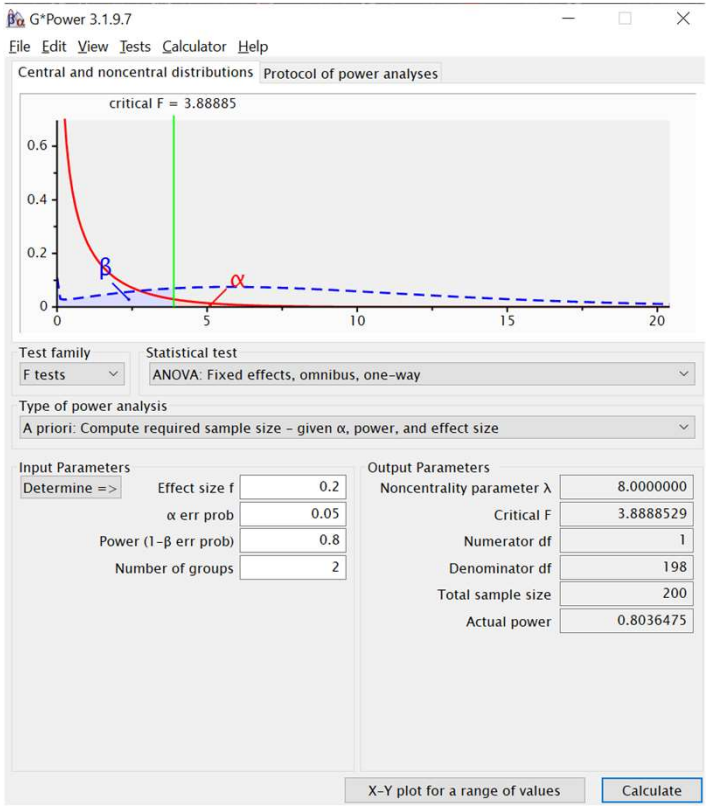
Analytical approach: Just look it up!

- Ideal for simple tests
- Fast
- Reliable (if you know where you’re looking)
- No coding required

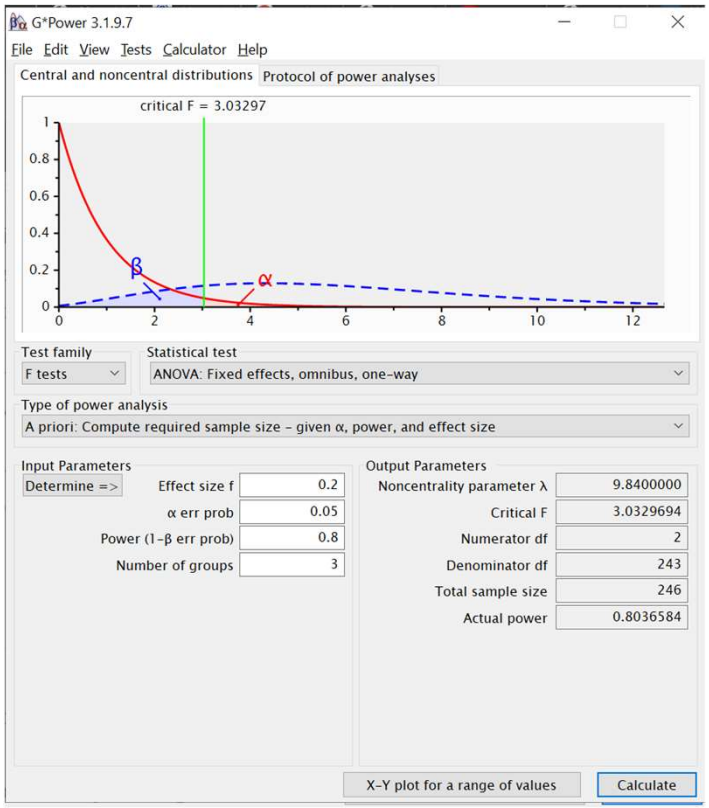
Risks of the analytical approach

“Statistical packages tend to be used as a kind of oracle …. In order to elicit a response from the oracle, one has to click one’s way through cascades of menus. After a magic button press, voluminous output tends to be produced that hides the [critical information] …, among lots of other numbers that are completely meaningless to the user, as befits a true oracle.” (Baayen, 2008)

One-way ANOVA with 2 groups: G*Power recommends 200 total participants



One-way ANOVA with 3 groups: G*Power recommends 246 total participants



Risks of the analytical approach

Additionally:

- Doesn’t tell you how many trials to run
- Doesn’t work with more complex designs
- Doesn’t work if your data isn’t normally distributed
- Doesn’t work if your group sizes are not of equal size

Generally: Doesn’t work if you are in an atypical situation of any kind

The alternative: Simulation!

- Model your data-generating process
- Use model to generate data
- Test whether we recover our model parameters

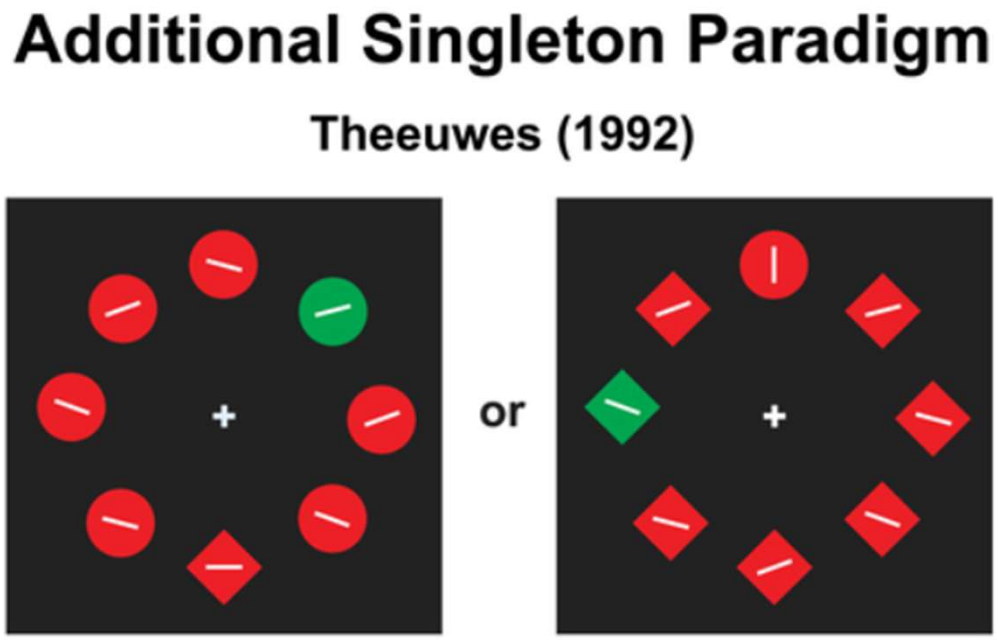
- Repeat 10,000 times!

The alternative: Simulation!

- (Almost) always works!
- Useful as a tool to aid thinking about your data
- Requires (some) coding
- Can take time

What we're doing next:

- Working through a toy example: Effect of singleton presence on visual search (e.g., Theeuwes 1992)



Our data-generating model:

- Population has a mean reaction time
- Manipulation (load) has a mean effect on reaction time
- Participants will vary from these means: random intercept and slope!
- Every trial also has some irreducible noise.

Our test

- Linear mixed-effects model
- Deal with irregular sample sizes, random effect of stimuli if desired
- Could move to glmm if considering e.g. non-linear outcomes (better model of RT, accuracy for recall, etc.)

Time for R!

Takeaways

- The importance of power for us as scientists
- Approaches to power analysis: the joys of simulation
- Beyond power analysis: simulation as a way to think critically about your study's assumptions and limitations!