Power analysis: bureaucratic busywork or critical part of the scientific method?

Quentin Read

USDA-ARS Southeast Area Statistician

- What is power analysis and why do we do it?
- What are effect sizes?
- There’s no free lunch in power analysis
- “Plug-and-chug” power analysis
- Power analysis by simulation for mixed models

What is statistical power?

What is power?

If a certain effect of interest exists (e.g. a difference between two groups) power is the chance that we actually find the effect in a given study.

Why do we do power analysis?

Why do we do power analysis?

~~Because we have to do it to get funding~~

Because it helps us design the study, foreseeing any issues that might arise

Because it reduces the chance of doing weak and inconclusive studies that are doomed before they even start

Because it elevates our science, and our science makes the world a better place

Why do we do power analysis?

We want to design studies that give us reliable answers to the questions we care about …

… minimizing time and resources spent …

… and minimizing harm to research subjects!

What you probably think power analysis is

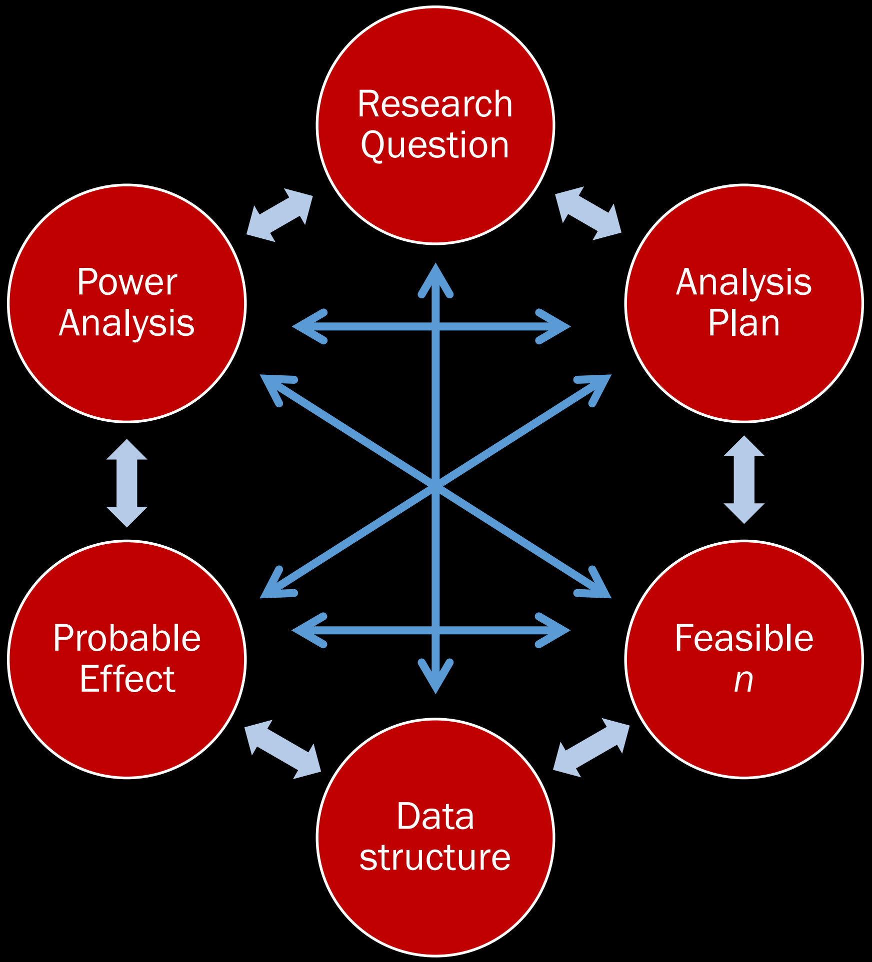
Research Question → Power Analysis → Number of participants

Slide borrowed from Jessica Logan (doi.org/10.6084/m9.figshare.8236409.v1)

What power analysis actually is

Iterative process

If any step doesn't work, you need to re-evaluate

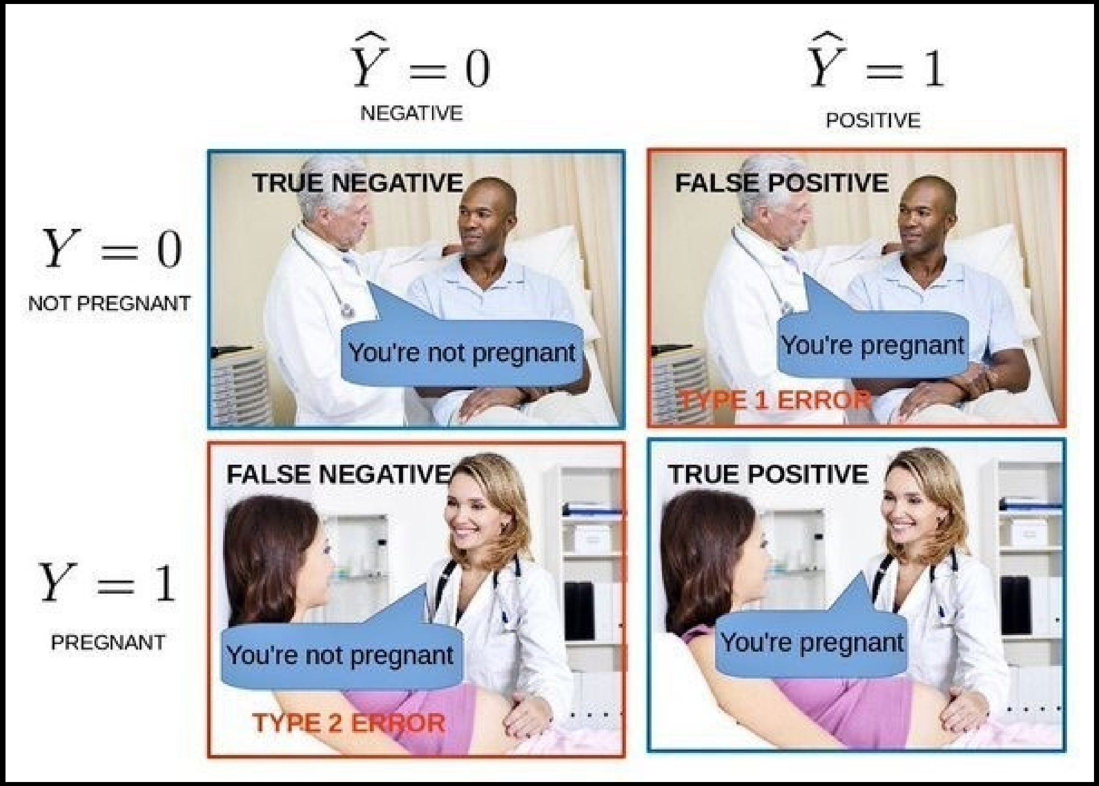


Slide borrowed from Jessica Logan (doi.org/10.6084/m9.figshare.8236409.v1)

False positives and false negatives

model prediction (Ŷ)

truth (Y)



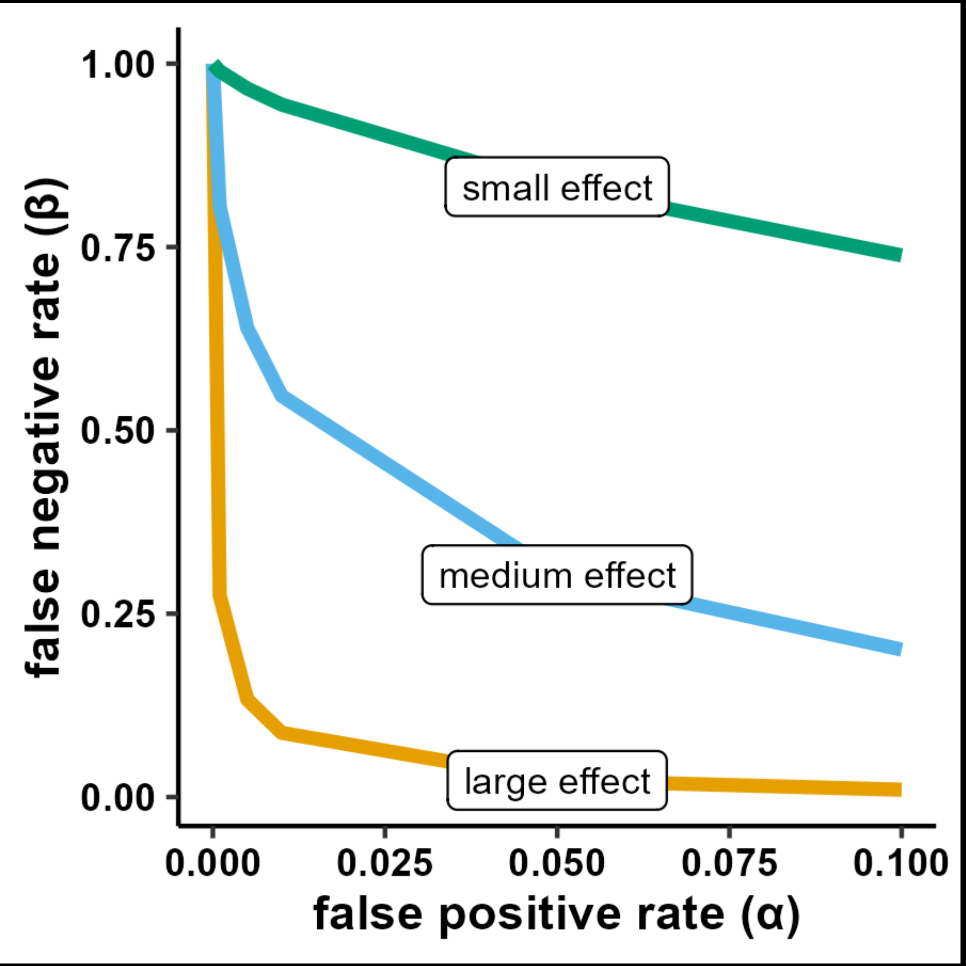
Often called “Type I and Type II error”

Conventional thresholds for power

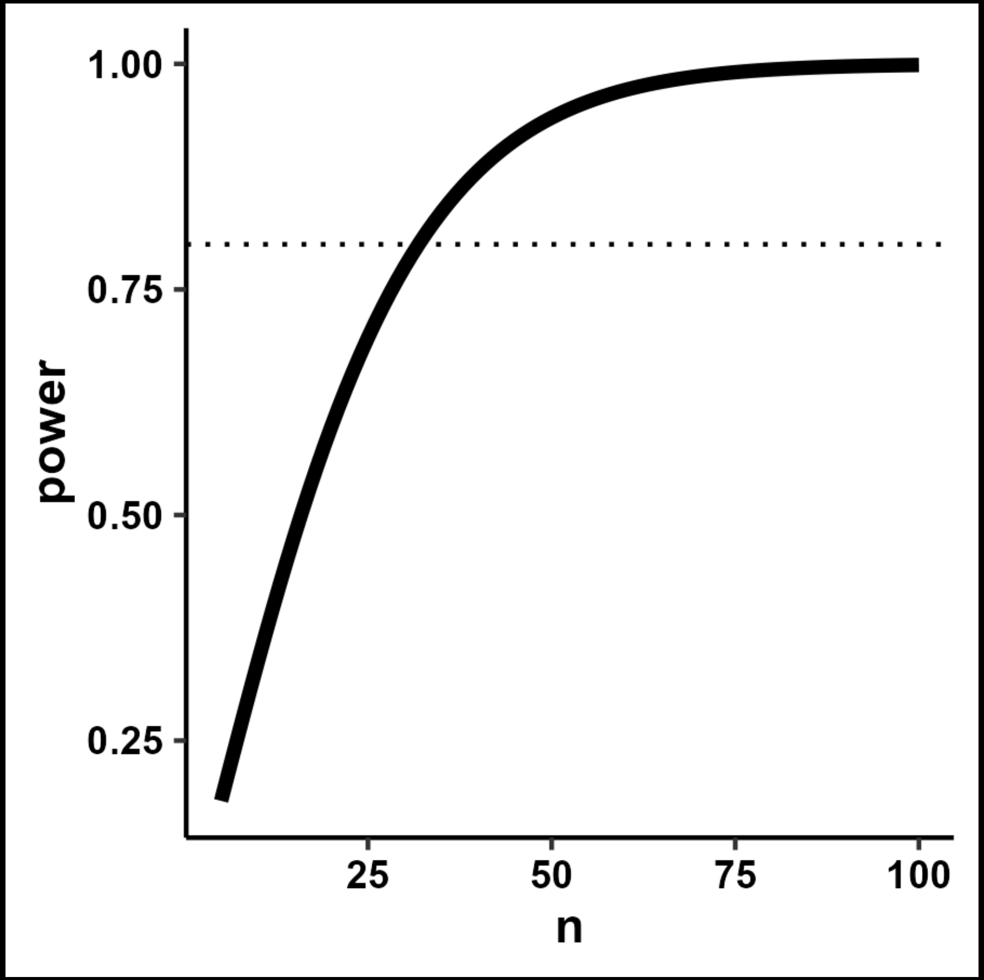
$\alpha = 0.05$, $\beta = 0.20$

4:1 ratio of risk of false negative : risk of false positive (conservative)

$1 - \beta = 1 - 0.2 = 0.8$ → “80% power”



Bigger sample size = more power … up to a point



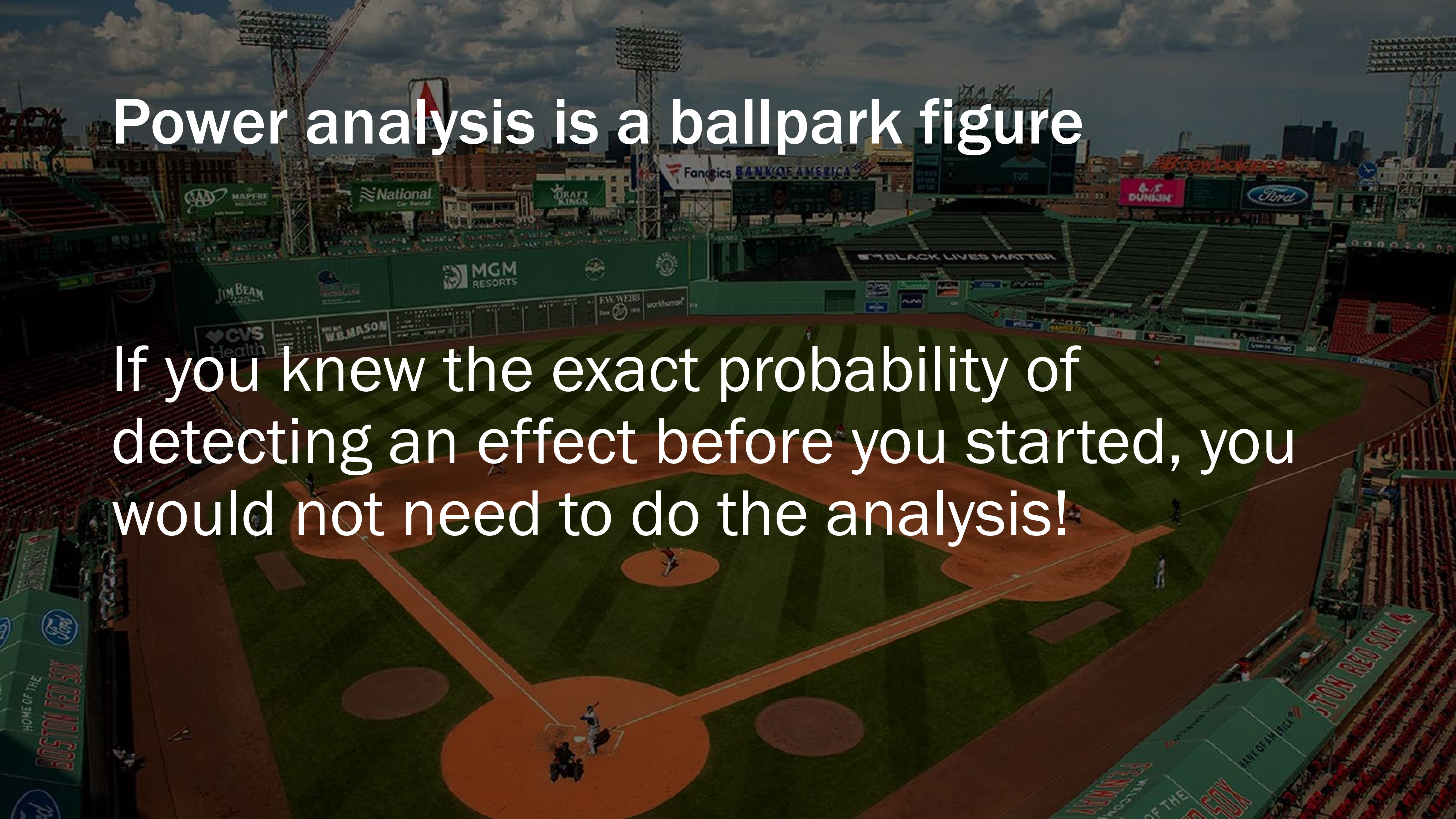
Power analysis should be conservative

Too many samples: excess resources wasted/animals harmed unnecessarily

"Just right": this number cannot be known in advance!

Too few samples: study is inconclusive. *ALL* resources wasted/animals harmed unnecessarily!

Err on the side of caution!



Power analysis is a ballpark figure

If you knew the exact probability of detecting an effect before you started, you would not need to do the analysis!

What do you need to know to do a power analysis?

Study design

Desired error rate for false positives and false negatives

Range of feasible sample sizes

Range of biologically relevant effect sizes

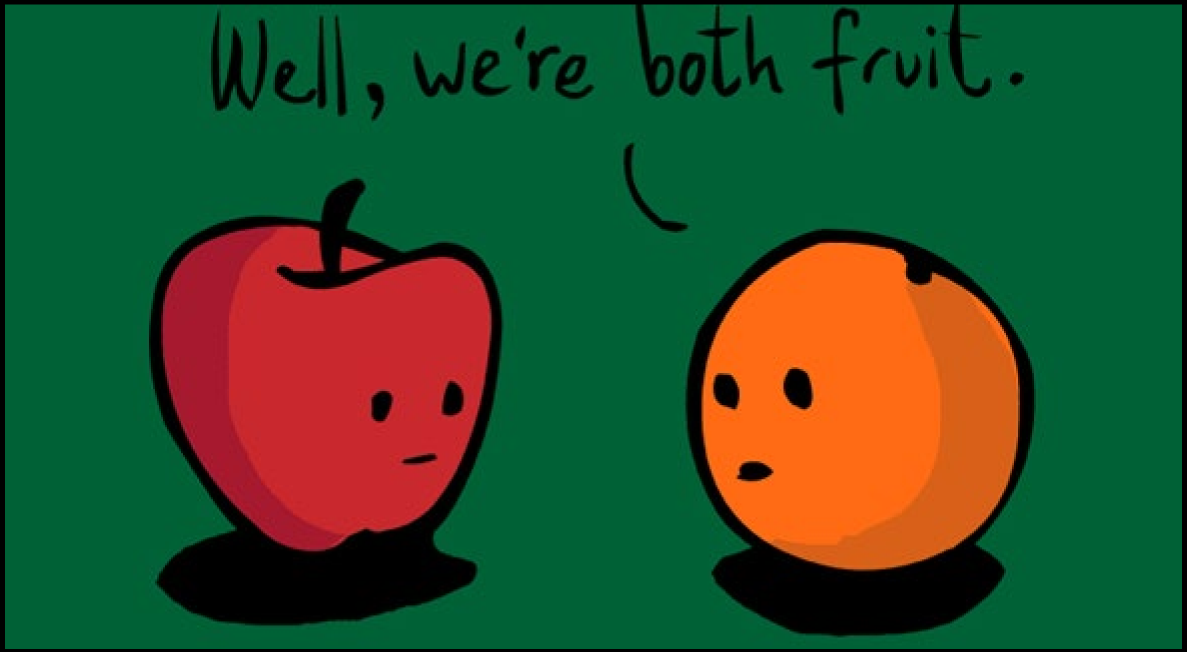
If you have ranges, repeat the power analysis for combinations of values across the ranges to generate a power curve

Standardized effect sizes

Biologically or clinically relevant effects: original units

"We want to be able to detect a body mass difference of 10 g between the two treatment means."

But effect sizes need to be comparable across studies and across measurement units!



Commonly used effect sizes

T-test design: Cohen's *d*

Difference between two means divided by pooled standard deviation. How many "standard deviation units" apart are two means?

ANOVA design: Cohen's *f*

d extended to >2 means; ratio of proportion of variance accounted for by group : proportion of variance *un*accounted for by group

Regression design: R^2 or f^2

f^2 : relative increase in R^2 when adding a predictor to a model

Effect size benchmarks

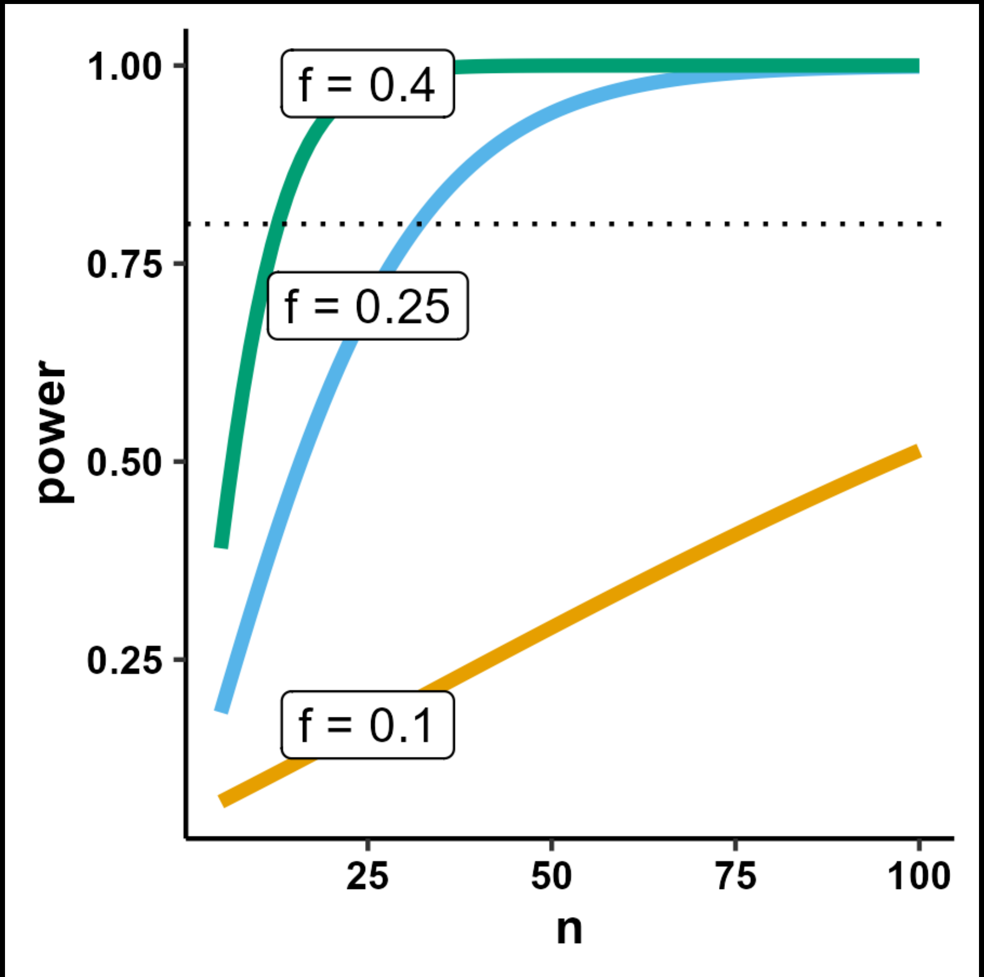
These numbers are used as fallbacks if no better knowledge exists

Not really appropriate to apply these benchmarks indiscriminately to data from all fields

This is a "lazy way out" and should only be used as a last resort

Cohen's d	Cohen's f	Strength of effect
0.2	0.1	Small
0.5	0.25	Medium
0.8	0.4	Large

Power depends on effect size



Knowns and unknowns

Scenario 1 We know sample size, rough estimate of effect size, and desired significance level. We calculate power of the study.

Scenario 2 We know rough estimate of effect size, desired significance level, and desired power. We can calculate sample size needed to get the desired power.

Scenario 3 We know sample size, desired significance level, and desired power. We can calculate the minimum detectable effect size (MDES).

What you put in → what you get out

You can do a power analysis just fine by making up plausible numbers

But the more background research you do the higher the quality of the power analysis … and the higher the quality of the resulting science

No shortcuts: It's hard work!!!

If you say "Do a power analysis for me" without providing the context, it may not be a good quality analysis

Where can we get background knowledge for power analysis effect size?

Your prior studies (preliminary studies or full-scale studies in a similar system)

Data from a literature review

This is why open data is important!

Fallback option: use a "small" or "medium" effect size

OPEN DATA

This is not just red tape!!!

Which effect size to use?

Do you care more about the F-test (ratio of variance within and between groups)?

Or is it specific difference(s) between means?

The latter is likely to be more clinically/biologically relevant

But it raises the issue of multiple comparisons

- Solution: Adjust the significance level (α) downward
- This requires higher sample size to get the same power

Plug-and-chug power analysis

For some very simple study designs, there are exact analytical formulas for statistical power

Use software to plug in parameters and get power estimates out

“Point and click” GUIs available, but scripting is better!

Plug-and-chug examples

Comparing the means of two groups with a t-test, using the R package **pwr**

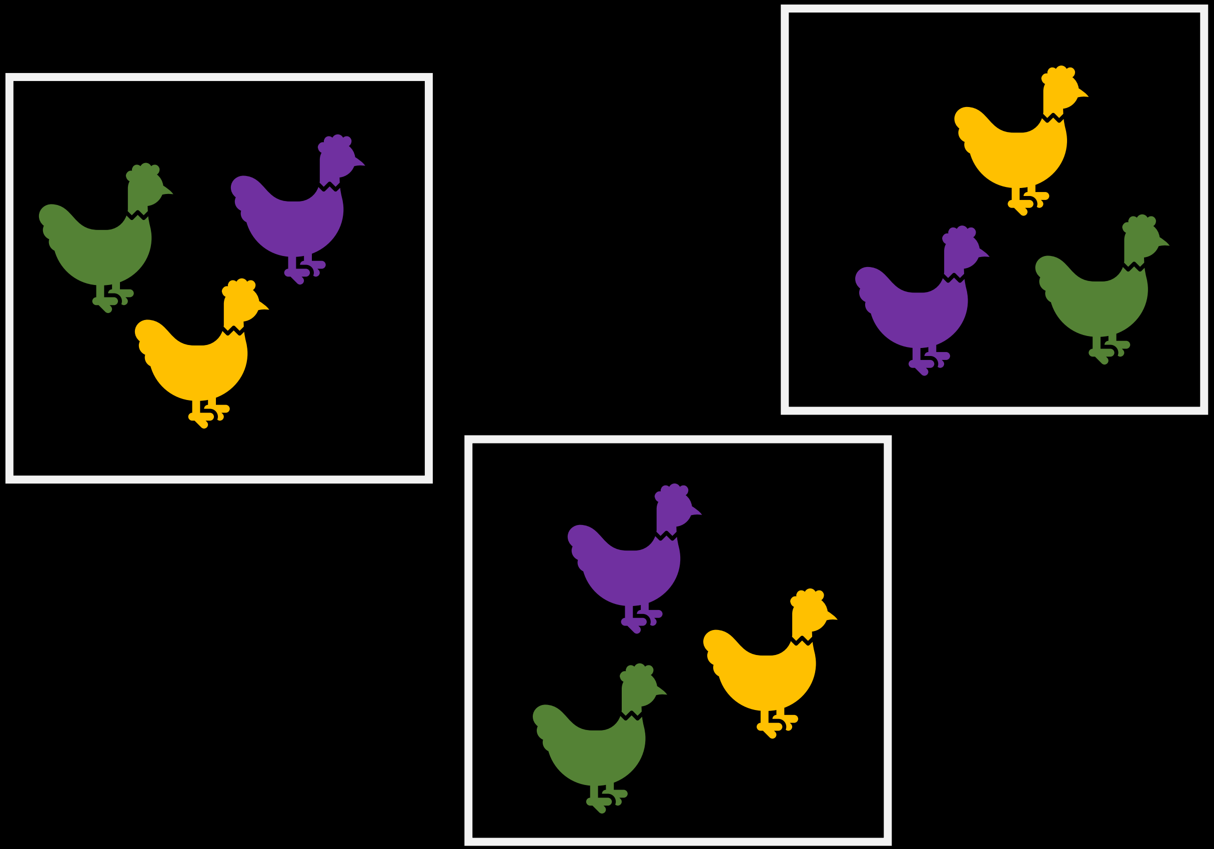
```
pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.8)
# Result: n = 64
```

We need 64 individuals per group to detect an effect size of 0.5 (Cohen's d) with 80% power

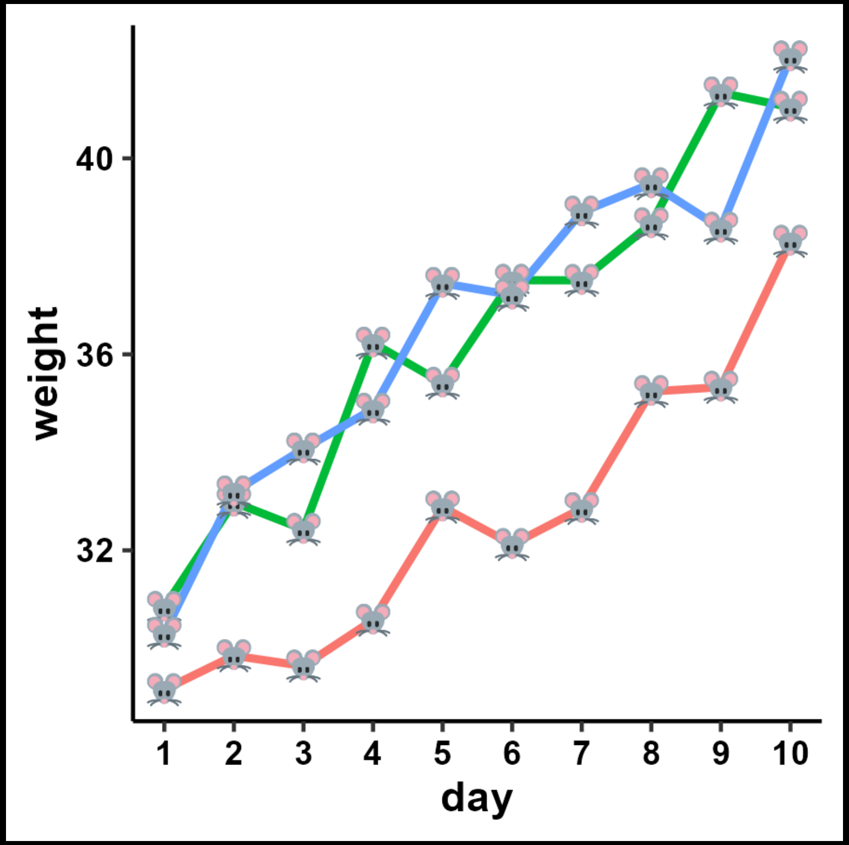
```
pwr.t.test(d = 0.75, sig.level = 0.05, n = 20)
# Result: power = 0.74
```

If we have 20 individuals per group, we have ~74% power to detect an effect size of 0.75

What about mixed models?



Randomized complete block designs



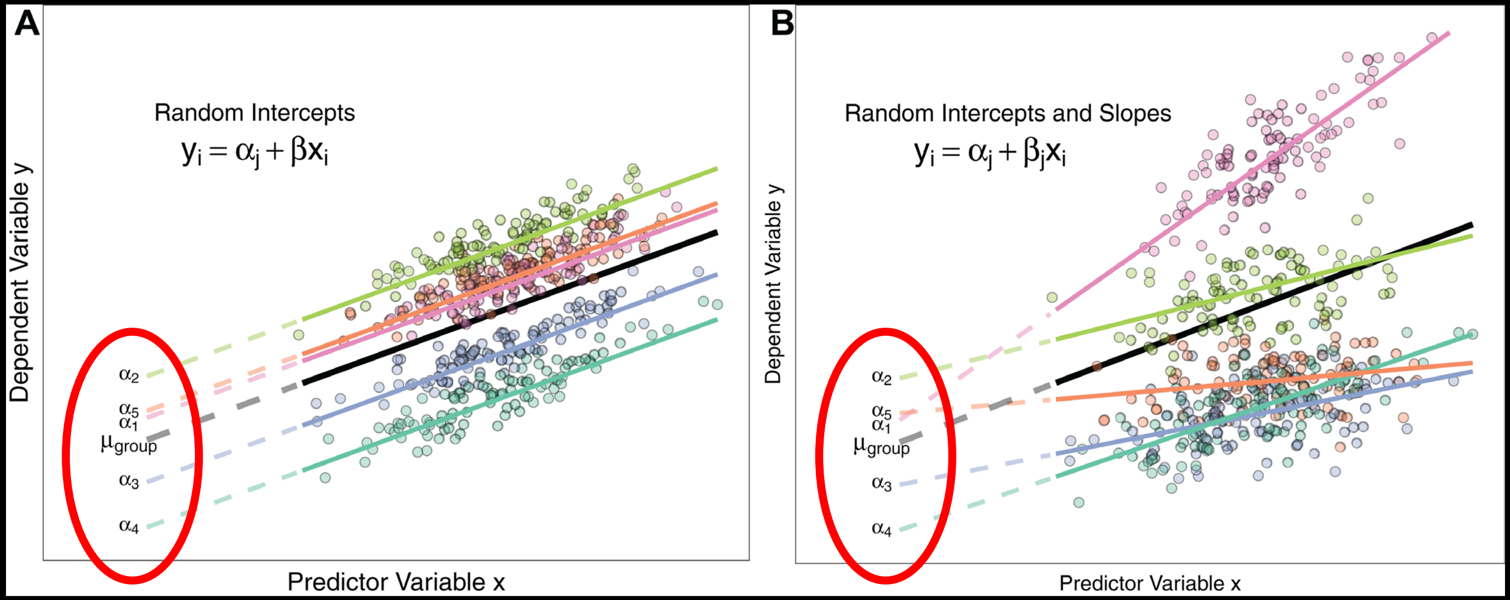
Repeated measures within individuals

Mixed model power analysis

To do a power analysis for a mixed model study design, we need to also have an estimate of how much random variation there is

Variance-covariance matrix of the random effects

This is not trivial!



Plug-and-chug vs. simulation

Most non-trivial experimental designs need power analysis by simulation

Mixed models

Bayesian analysis

If more informed, the power analysis will be better

Simulation makes it easier to test power to detect effects on the raw scale … maybe more intuitive?

We usually construct power curves, where we can look at power results for a variety of different assumptions

As always, it’s important to be conservative

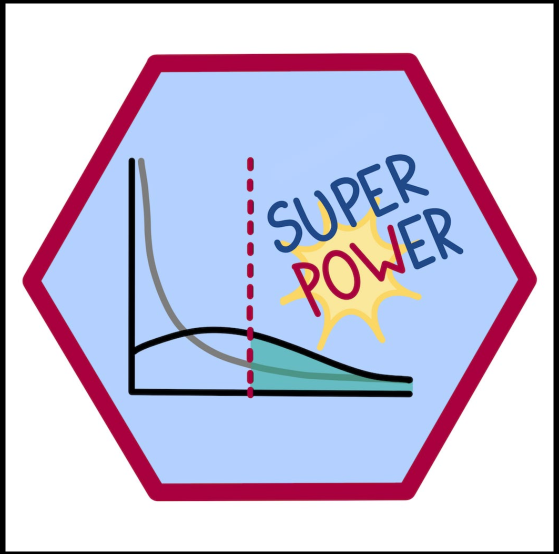
Software implementations

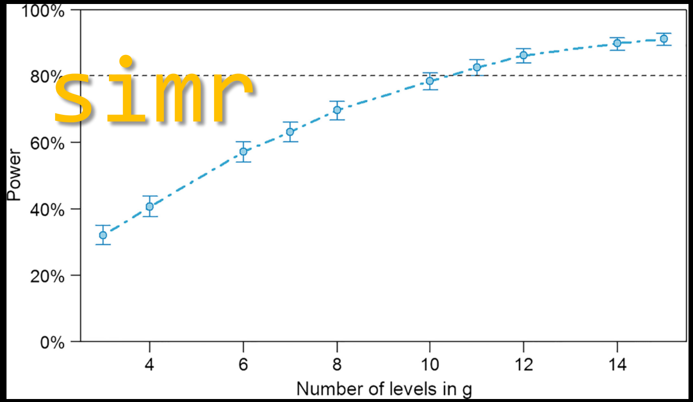
G*Power

R packages `SuperPower`, `WebPower`, `pwr`

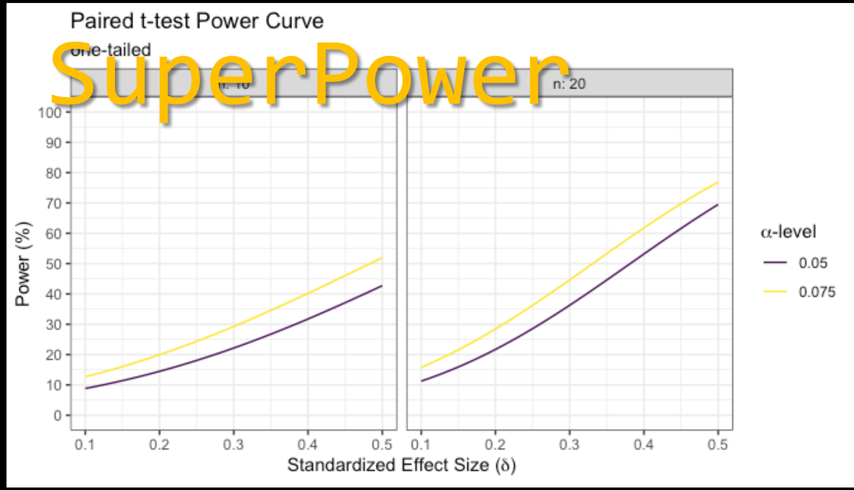
R packages `paramtest`, `simr`

SAS: `proc power`, `proc glmpower`





Green & MacLeod 2015, *Methods Ecol. Evol.*



Lakens & Caldwell 2022

Type S and Type M error

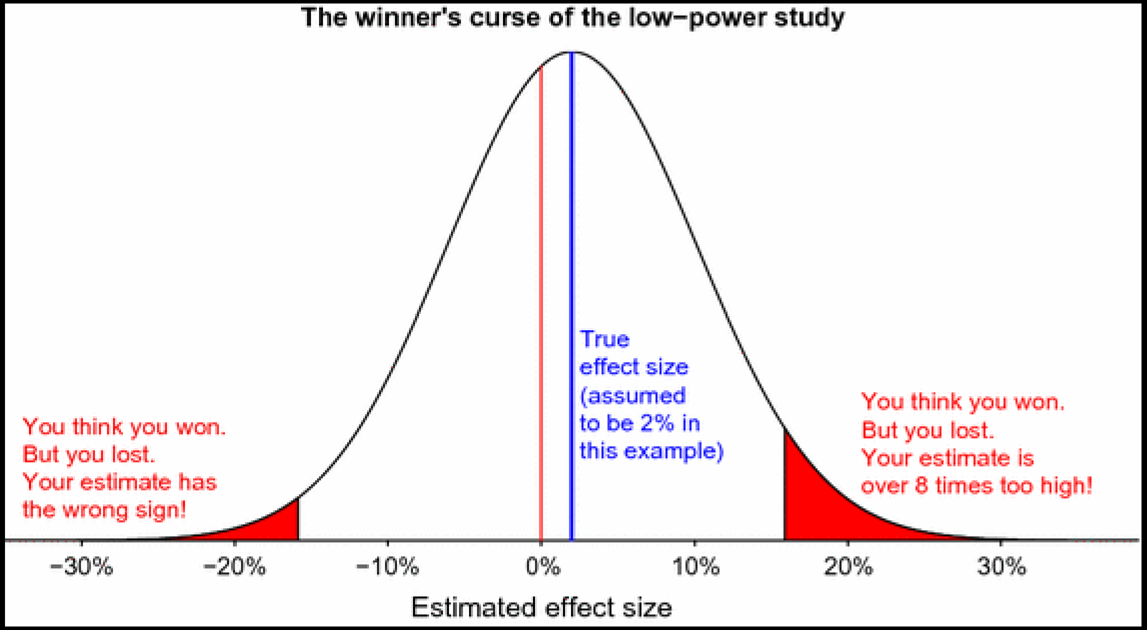
A new paradigm: Gelman et al. critique the focus on black and white "false + vs. false –"

Type S: the sign is wrong

Type M: the magnitude is wrong

You can be "wrong" even if you say there is an effect when there truly is one, if the magnitude is wrong

Even more impetus to design appropriately powered studies



Gelman, Skardhamar, & Aaltonen 2017, *J. Quant. Criminology*

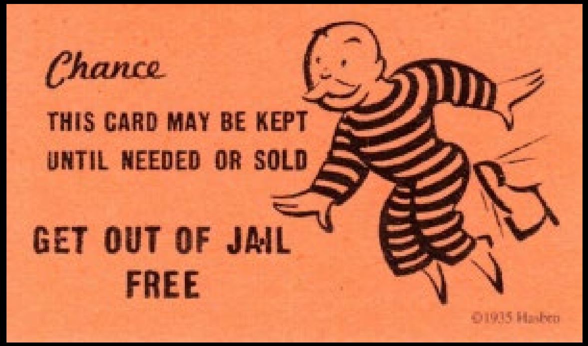
What did we learn today?

Closing remarks

The better informed your power analysis is by literature review and preliminary data, the better outcome you will get …

… both in the power analysis and in the resulting science!

Do low-powered studies at your own risk! There's no "get out of jail free card"



If you aren't sure, talk to your friendly neighborhood statistician!

usda-ree-ars.github.io/SEAStats
quentinread.com
quentin.read@usda.gov