

Do we need more power?

Brice Ozenne

December 8, 2017, NRU Christmas Symposium

Setting

We consider:

- an outcome Y e.g. fMRI
- an exposure variable E e.g. SAD, season

We would like to compare:

- μ_0 the expected outcome under exposure E_0
- μ_1 the expected outcome under exposure E_1

i.e. to test the null hypothesis:

$$(\mathcal{H}_0) \mu_0 = \mu_1 \quad \text{vs.} \quad (\mathcal{H}_1) \mu_0 \neq \mu_1$$

with a risk of false positive of $\alpha = \mathbb{P}[\text{reject } \mathcal{H}_0 | \mathcal{H}_0 \text{ is true}] = 0.05$.

Reporting power

Reviewer: Power should be reported [...] this information is vital for interpretation

Reporting power

Reviewer: Power should be reported [...] this information is vital for interpretation

Me: 🙄 Should I buy new statistical textbooks ?????

About power

The power of a statistical test is defined by:

$$\begin{aligned}\beta &= 1 - \mathbb{P}[\text{reject } \mathcal{H}_0 | \mathcal{H}_1 \text{ is true}] \\ &= f\left(\alpha, \mu_0 - \mu_1, \sigma_{\hat{\mu}_0 - \hat{\mu}_1}\right)\end{aligned}$$

$\hat{\sigma}_{\hat{\mu}_0 - \hat{\mu}_1}$ is the uncertainty about the estimate:

- simple settings: $\sigma_{\hat{\mu}_0 - \hat{\mu}_1} = \frac{\sigma(Y)}{\sqrt{n}}$

So we could estimate the power if we would know $\mu_0 - \mu_1$.

- but we don't ! (it is the aim of the study)

Practical solutions to compute the power (1/2)

What about using the estimated effects $\widehat{\mu}_0$ and $\widehat{\mu}_1$?

- called observed power
- is uninformative
- and often wrong
(i.e. biased when conditioned on
keeping/rejecting \mathcal{H}_0)

Practical solutions to compute the power (1/2)

What about using the estimated effects $\widehat{\mu}_0$ and $\widehat{\mu}_1$?

- called observed power
- is uninformative
- and often wrong
(i.e. biased when conditioned on keeping/rejecting \mathcal{H}_0)

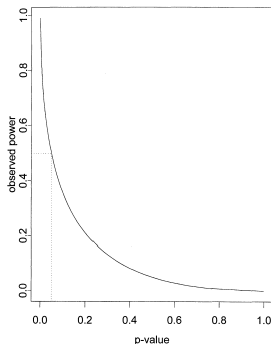


Figure 1. "Observed" Power as a Function of the p Value for a One-Tailed Z Test in Which α is Set to .05. When a test is marginally significant ($P = .05$) the estimated power is 50%.

(?)

Practical solutions to compute the power (2/2)

What about using a priori knowledge of $\widehat{\mu}_0$ and $\widehat{\mu}_1$?

- if we are guessing $\widehat{\mu}_0$ and $\widehat{\mu}_1$:
 - power can be informative when planning experiments ...
 - but should not be taken too seriously.
- if we want to replicate a study:
 - power is very convoluted way to compare two studies

Limitations of studies with sample size

- (i) Lower representativity of the population of interest
- (ii) Larger uncertainty on the estimates
- (iii) Increased type 2 error

Limitations of studies with sample size

- (i) Lower representativity of the population of interest
- (ii) Larger uncertainty on the estimates
- (iii) Increased type 2 error

Reviewer: [...] underpowered analyses that carry higher risk for false positives.

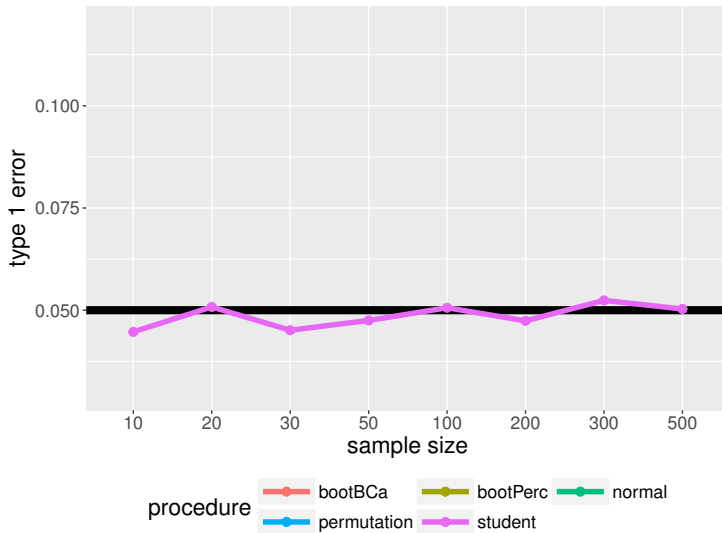
Limitations of studies with sample size

- (i) Lower representativity of the population of interest
- (ii) Larger uncertainty on the estimates
- (iii) Increased type 2 error

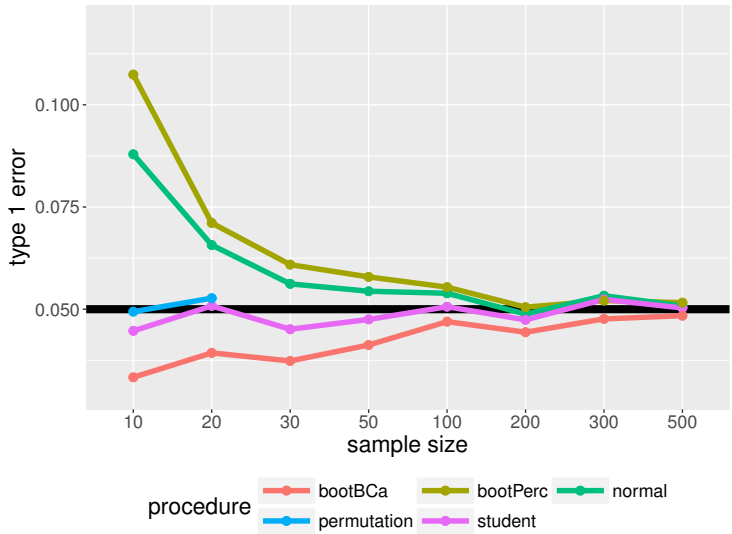
Reviewer: [...] underpowered analyses that carry higher risk for false positives.

Me: 🙄 definition of underpowered study?
Do you mean with small sample size ...

False positive: the sample size can matter



False positive: the sample size can matter



Example: adjusting for multiple comparisons

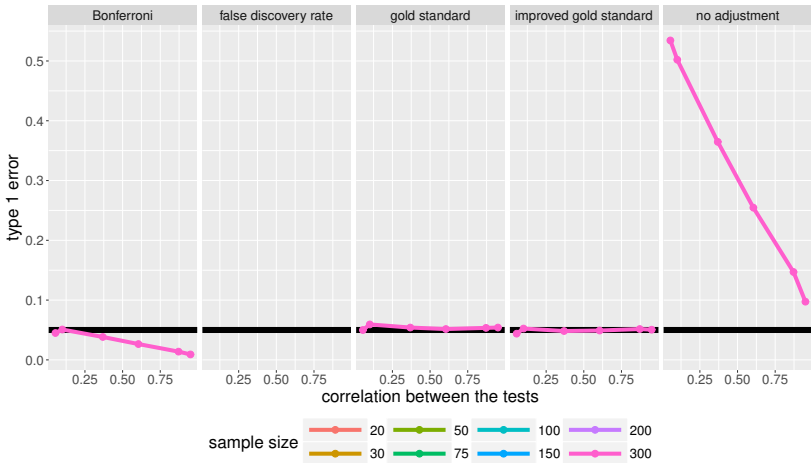
Relationship between:

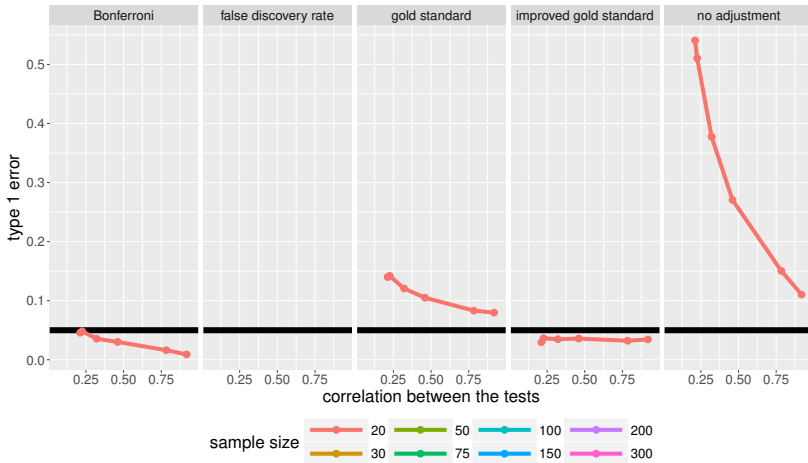
- treatment
- several psychological (correlated) outcomes ($m=15$)

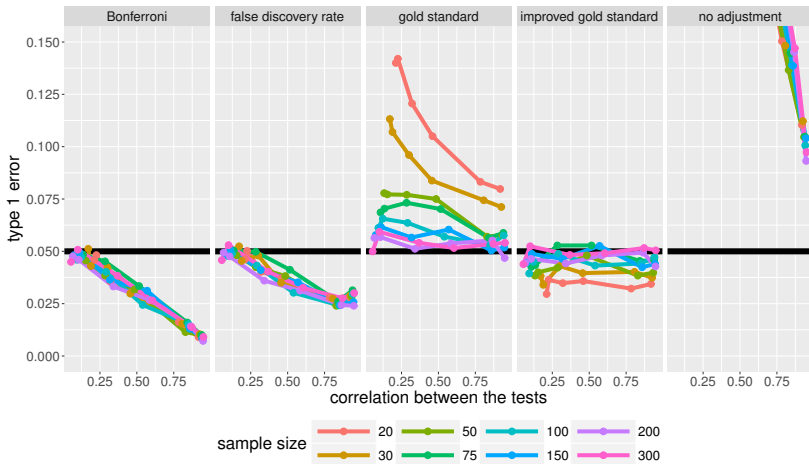
Suppose we fit a (linear) model specific to each outcome.

We would like to report a confidence interval / p.value for the most significant outcome.

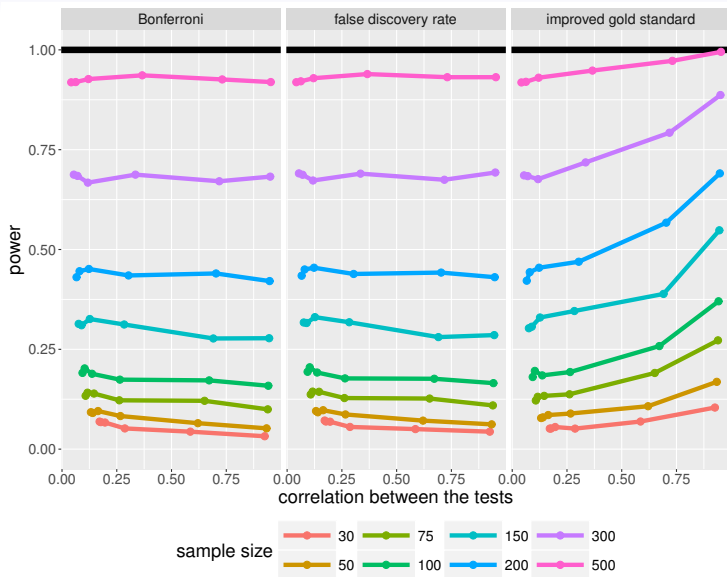
- keep a risk of false positive of $\alpha = 0.05$







Note: Comparison with false discovery rate is not fair



Another argument

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button^{1,2}, John P. A. Ioannidis³, Claire Mokrysz¹, Brian A. Nosek⁴, Jonathan Flint⁵, Emma S. J. Robinson⁶ and Marcus R. Munafò¹

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

Diagnostic tests

Consider

- a disease D e.g. depression
- a diagnostic test D e.g. MDI
- an existing knowledge prevalence of depression

What is the probability of having the disease given a positive diagnostic test?

$$\mathbb{P}[D = 1 | T = 1] = \frac{\mathbb{P}[T = 1 | D = 1] \mathbb{P}[D = 1]}{\mathbb{P}[T = 1]}$$

- We (implicitly) define a population of interest (e.g. danish citizen in 2012)
- Bayesian approach: update existing knowledge

Their argument

What is the probability of \mathcal{H}_0 to be false when rejecting \mathcal{H}_0 :


$$\mathbb{P}[\mathcal{H}_0 \text{ false} | \text{reject } \mathcal{H}_0] = \frac{(1 - \beta) * R}{(1 - \beta) * R + \alpha}$$

R: "the odds that a probed effect is indeed non-null among the effects being probed"


Comments:

- They don't define the population of interest.
- Bayesian approach: what is our prior knowledge?
- Is that wise to reduce a study to a binary decision?

Conclusion

Power  hazard

Be careful

α  β

?? $PPV = ([1 - \beta] \times R) / ([1 - \beta] \times R + \alpha)$??

References I

Publication bias

"consequences [...] include overestimates of effect size and low reproducibility of results ..."



Publication bias

"consequences [...] include overestimates of effect size and low reproducibility of results ..."

