Effect size and power

Session 5

MATH 80667A: Experimental Design and Statistical Methods
for Quantitative Research in Management
HEC Montréal

Outline

Effect sizes

Power

Effect size

Example from the OSC psychology replication

> The key statistics provided in the paper to test the "depletion" hypothesis is the main effect of a one-way ANOVA with three experimental conditions and confirmatory information processing as the dependent variable ($F(2, 82) = 4.05, p = 0.02, \eta^2 = 0.09$). Considering the original effect size and an alpha of 0.05 the sample size needed to achieve 90% power is 132 subjects.

Replication report of Fischer, Greitemeyer, and Frey (2008, JPSP, Study 2) by E.M. Galliani

Q: What is the sample size for given power?

Q: How big is this effect?

Does it matter?

Statistical significance $\neq$ practical relevance

With large enough sample size, **any** sized difference between treatments becomes statistically significant.

But whether this is important depends on the scientific question.

Example

- What is the minimum difference between two treatments that would be large enough to justify commercialization of a drug?
- Tradeoff between efficacy of new treatment vs status quo, cost of drug, etc.

Measures of effects

F-statistics and p-values are not good summaries of effect size:

- the larger the sample size, the bigger the statistic

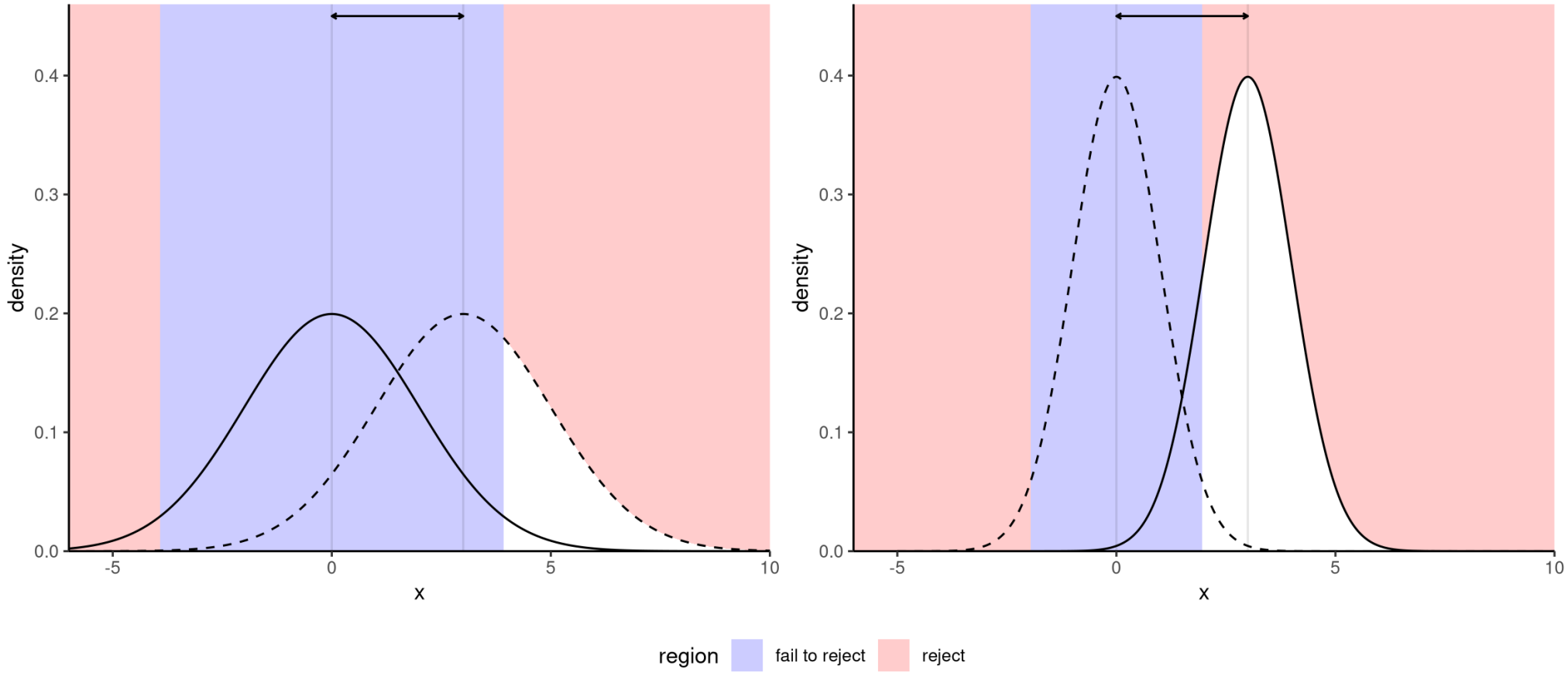
Instead use

- standardized differences/measures
- percentage of variability explained

Popularized in the handbook

> Cohen, Jacob. Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Routhledge, 1988.

Illustrating effect size (differences)



The plot shows null (thick) and true distributions (dashed) for sample mean with small (left) and large (right) samples. The effect size (distance between means) is the same.

Estimands, estimators, estimates

- μ_i is the (unknown) population mean of group i (parameter, or estimand)
- $\widehat{\mu}_i$ is a formula (an estimator) that takes data as input and returns a numerical value (an estimate).
- throughout, use hats to denote estimated quantities:

Left to right: parameter μ (target), estimator $\widehat{\mu}$ (recipe) and estimate $\widehat{\mu} = 10$ (numerical value, proxy)

Cohen's *d*

Standardized measure of effect (dimensionless=no units):

Assuming equal variance σ^2, compare mean of two groups i and j:

$$d = \frac{\mu_i - \mu_j}{\sigma}$$

The usual estimator $\hat{d}$ uses sample average of groups and the pooled variance $\widehat{\sigma}$. Note: a finite sample correction (Hedge) can be used.

Cohen's *d* is sometimes reported in terms of effect size

- small (*d=0.2*), medium (*d=0.5*) or large (*d=0.8*).

Cohen's f

With more than two groups and assuming equal variance σ^2, compare the squared difference between overall mean and group mean

$$f^2 = \frac{1}{\sigma^2} \sum_{j=1}^{k} \frac{n_j}{n} (\mu_j - \mu)^2,$$

a weighted sum of squared difference relative to the overall mean μ.

For $k = 2$ groups, $f = d/2$.

Effect size: proportion of variance

Break down the variability $\sigma^2_{\text{total}} = \sigma^2_{\text{resid}} + \sigma^2_{\text{effect}}$ and define the percentage of variability explained by the effect.

$$\eta^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{\sigma^2_{\text{effect}}}{\sigma^2_{\text{total}}}$$

Often, you see instead the partial value

$$\eta^2_p = \frac{\sigma^2_{\text{effect}}}{\sigma^2_{\text{error}} + \sigma^2_{\text{effect}}}.$$

Note: for a one-way ANOVA (no repeated measurements), the two are equivalent.

Coefficient of determination

For the balanced one-way ANOVA, typical estimator is

$$\widehat{R}^2 = \frac{F\nu_1}{F\nu_1 + \nu_2}$$

where $\nu_1 = k - 1$ and $\nu_2 = n - k$ are the degrees of freedom for the one-way ANOVA.

- $\widehat{R}^2$ is an upward biased estimator (too large).
- People frequently write η^2 when they mean $\widehat{R}^2$
- for the replication, $\widehat{R}^2 = (4.05 \times 2)/(4.05 \times 2 + 82) = 0.09$

ω^2 square

Another estimator of η^2 that is recommended in Keppel & Wickens (2004) for power calculations is $\widehat{\omega}^2$.

For one-way ANOVA, the latter is obtained from the F-statistic as

$$\widehat{\omega}^2 = \frac{\nu_1(F-1)}{\nu_1(F-1)+n}$$

- for the replication, $\widehat{\omega}^2 = \frac{3.05\times 2}{3.05\times 2+84} = 0.0677.$
- if the value returned is negative, report zero.

Converting η^2 to Cohen's f

The software we will use take an estimate of Cohen's f (or f^2) as input for the effect size.

Convert from one to the other:

$$f^2 = \frac{\eta^2}{1 - \eta^2}.$$

If we plug-in estimated values $\widehat{R}^2$ and $\widehat{\omega}^2$, we get $\widehat{f} = 0.314$ and $\tilde{f} = 0.27$.

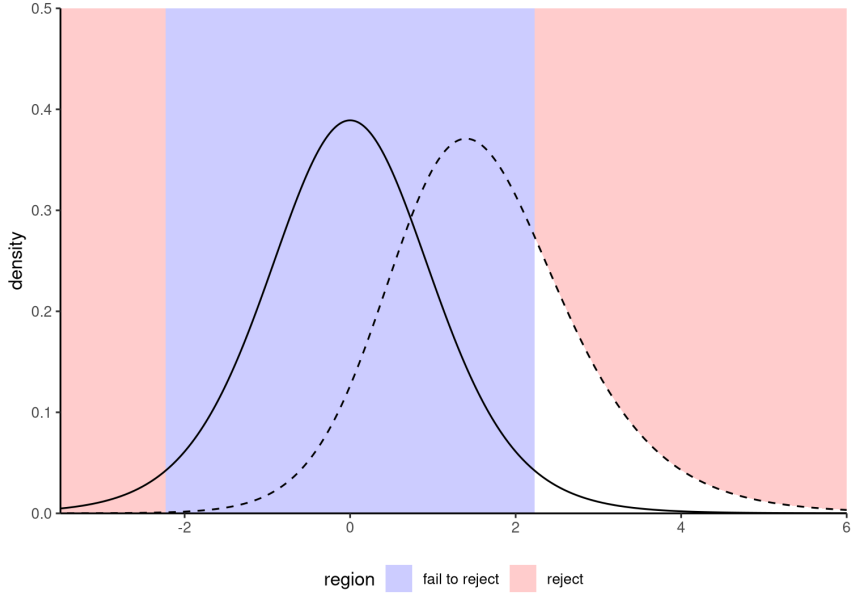
Comments on effect estimates

- There are two variants: population quantities (e.g., η^2) that depend on unknown parameters and sample estimates (e.g., $\widehat{R}^2$, $\widehat{\omega}^2$)
- In more complicated models, we can look at partial effects (proportion of variance relative to that of errors)
- Every effect size estimator is random (because its inputs are): **hugely uncertain**
- We can report confidence intervals with estimates (mostly for Cohen's d, but these are hopelessly wide in most settings).

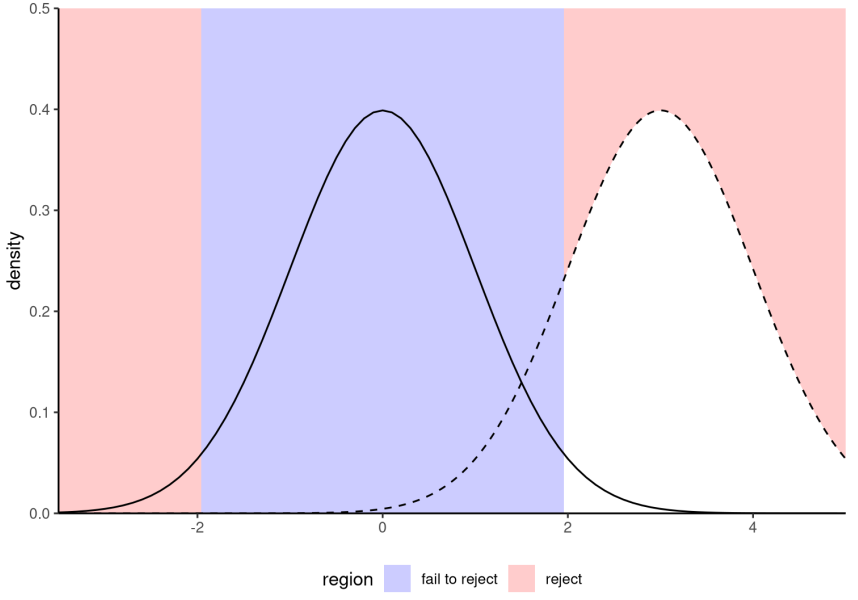
Power

I cried power!

The null alternative corresponds to a single value (equality in mean), whereas there are infinitely many alternatives...



Power is the ability to detect when the null is false, for a given alternative (dashed).



Power is the area in white under the dashed curved, beyond the cutoff.

Parametrization of one-way ANOVA

group j has n_j observations

population average of group j is μ_j

We can parametrize the model in terms of the overall sample average,

$$\mu = \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\mu_j = \frac{1}{n}\sum_{j=1}^{K} n_j\mu_j,$$

where $n = n_1 + \cdots + n_K$ is the total sample size.

What determines power?

Think in your head of potential factors.

1. The size of the effects, $\delta_1 = \mu_1 - \mu, \ldots, \delta_K = \mu_K - \mu$
2. The background noise (intrinsic variability, σ^2)
3. The level of the test, α
4. The sample size in each group, n_j
5. The choice of experimental design
6. The choice of test statistic

We focus on the interplay between

effect size | **power** | **sample size**

Power and sample size calculations

Journals and grant agencies oftentimes require an estimate of the sample size needed for a study.

- large enough to pick-up effects of scientific interest (good signal-to-noise)
- efficient allocation of resources (don't waste time/money)

Same for replication studies: how many participants needed?

Living in an alternative world

Recall that with K treatments (groups) n observations, the F-statistic is

$$F = \frac{\text{between sum of squares}/(K-1)}{\text{within sum of squares}/(n-K)}$$

The null distribution is $F(K-1, n-K)$.

The denominator is an estimator of σ^2 under both the null and alternative.

So how does the *F*-test behaves under an alternative?

Numerator of the *F*-test

What happens to the numerator?

$$\mathsf{E}(\text{between sum of squares}) = \sigma^2\{(K-1) + \Delta\}.$$

where

$$\Delta = \frac{\sum_{j=1}^{K} n_j(\mu_j - \mu)^2}{\sigma^2} = nf^2.$$

Under the null hypothesis, $\mu_j = \mu$ for $j = 1, \ldots, K$ and $\Delta = 0$.

The greater Δ, the further the mode (peak of the distribution) is from zero.

Noncentrality parameter and power

$$\Delta = \frac{\sum_{j=1}^{K} n_j(\mu_j - \mu)^2}{\sigma^2}.$$

When does power increase?

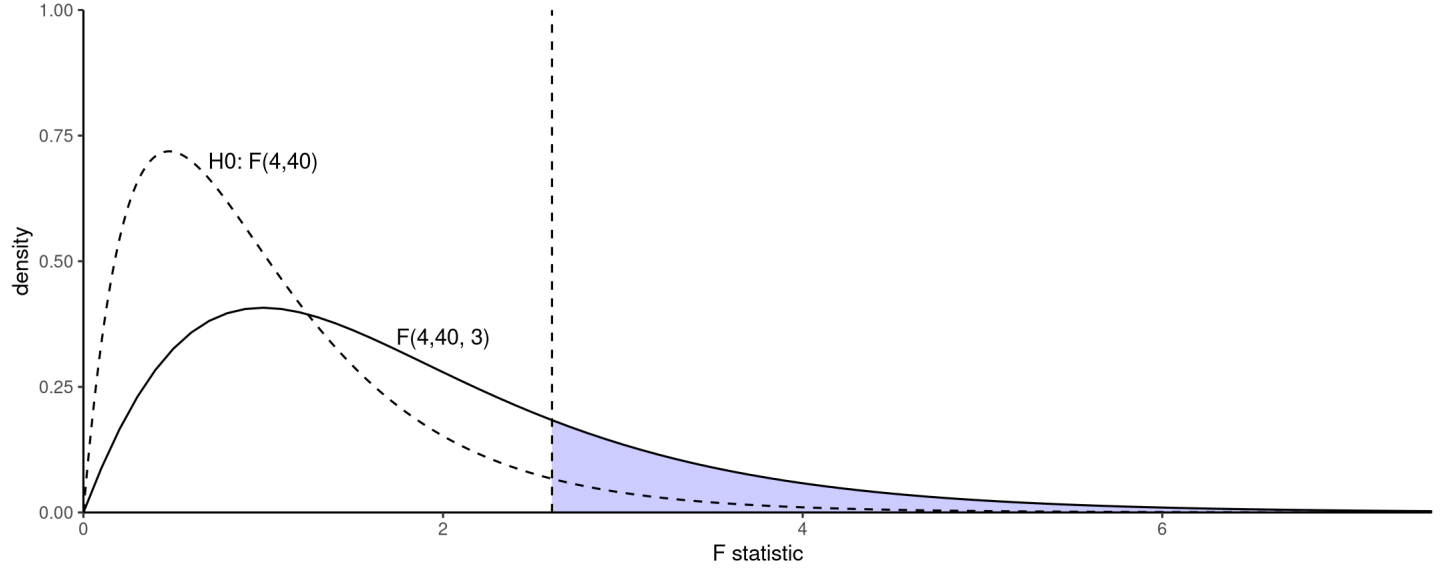
What is the effect of an increase of the

- group sample size $n_1, \ldots, n_K$.
- variability σ^2.
- true mean difference $\mu_j - \mu$.

Noncentrality parameter

The distribution is $F(\nu_1, \nu_2, \Delta)$ distribution with degrees of freedom ν_1 and ν_2 and noncentrality parameter Δ.

One-way ANOVA with n observations and K groups: $\nu_1 = K - 1$ and $\nu_2 = n - K$.



Note: the $F(\nu_1, \nu_2)$ distribution is indistinguishable from $\chi^2(\nu_1)$ for ν_2 large.

Computing power

Given a value of $\Delta = nf^2$, we can compute the tail probability as follows

1. Compute the cutoff point: the value under $\mathscr{H}_0$ that leads to rejection at level α.

```
cutoff <- qf(p = 1-alpha, df1 = df1, df2 = df2)
```

2. Compute probability below the alternative curve, from the cutoff onwards.

```
pf(q = cutoff, df1 = df1, df2 = df2, ncp = Delta, lower.tail =
                                                    FALSE)
```

How do we compute the power

Assume that the design is balanced, meaning $n_1 = \cdots = n_k = n/k$.

Then,

$$\Delta = \frac{n}{k\sigma^2} \sum_{j=1}^{k} (\mu_j - \mu)^2.$$

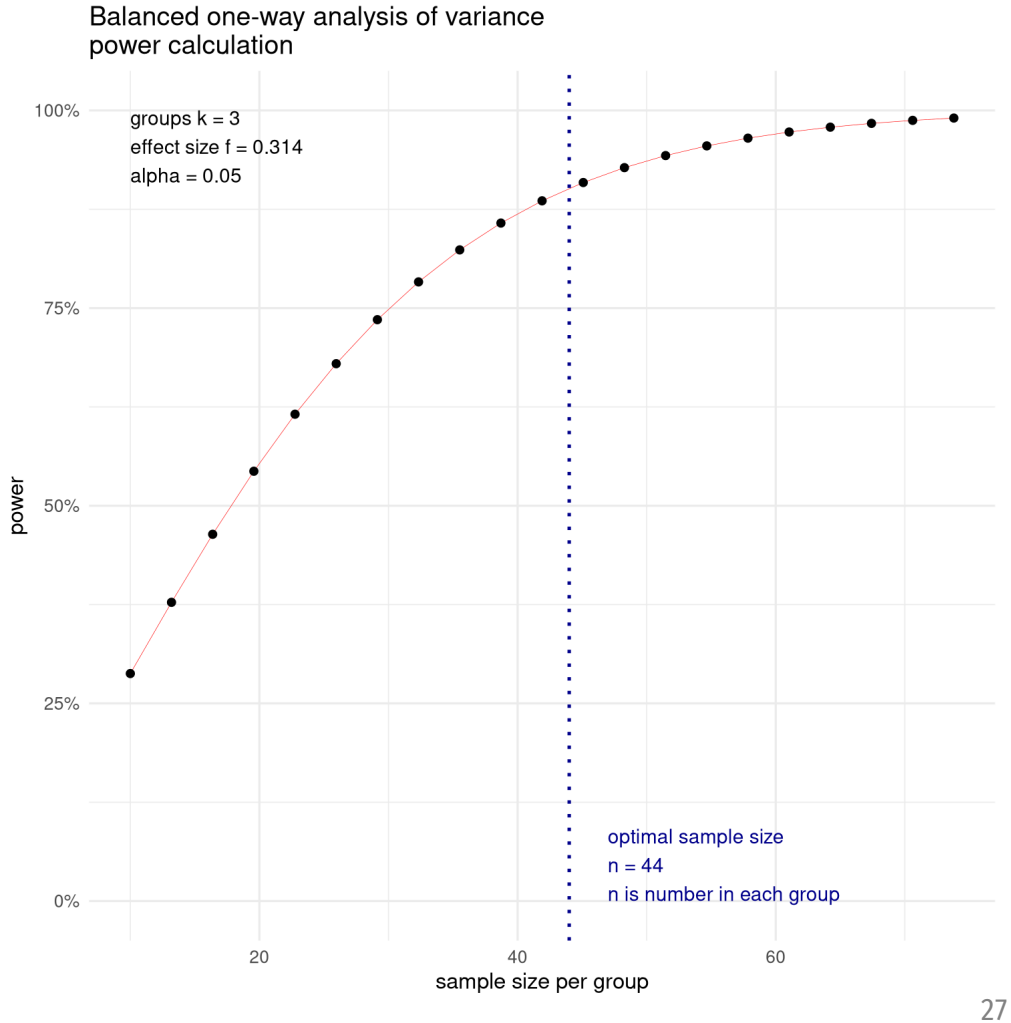
Plug-in $\texttt{df1} = k - 1$, $\texttt{df2} = n - k$ and $\texttt{ncp} = \Delta$ for fixed mean difference, level and number of groups in the formulas of the previous slide.

Power curves

```
library(pwr)
power_curve <-
 pwr.anova.test(
  f = 0.314, #from R-squared
  k = 3,
  power = 0.9,
  sig.level = 0.05)
plot(power_curve)
```

Recall: convert η^2 to Cohen's f (the effect size reported in `pwr`) via $f^2 = \eta^2/(1-\eta^2)$

Using $\tilde{f}$ instead (from $\widehat{\omega}^2$) yields $n = 59$ observations per group!



Effect size estimates

WARNING!

Most effects reported in the literature are severely inflated.

Publication bias & the file drawer problem

Estimates reported in meta-analysis, etc. are not reliable. Use scientific knowledge

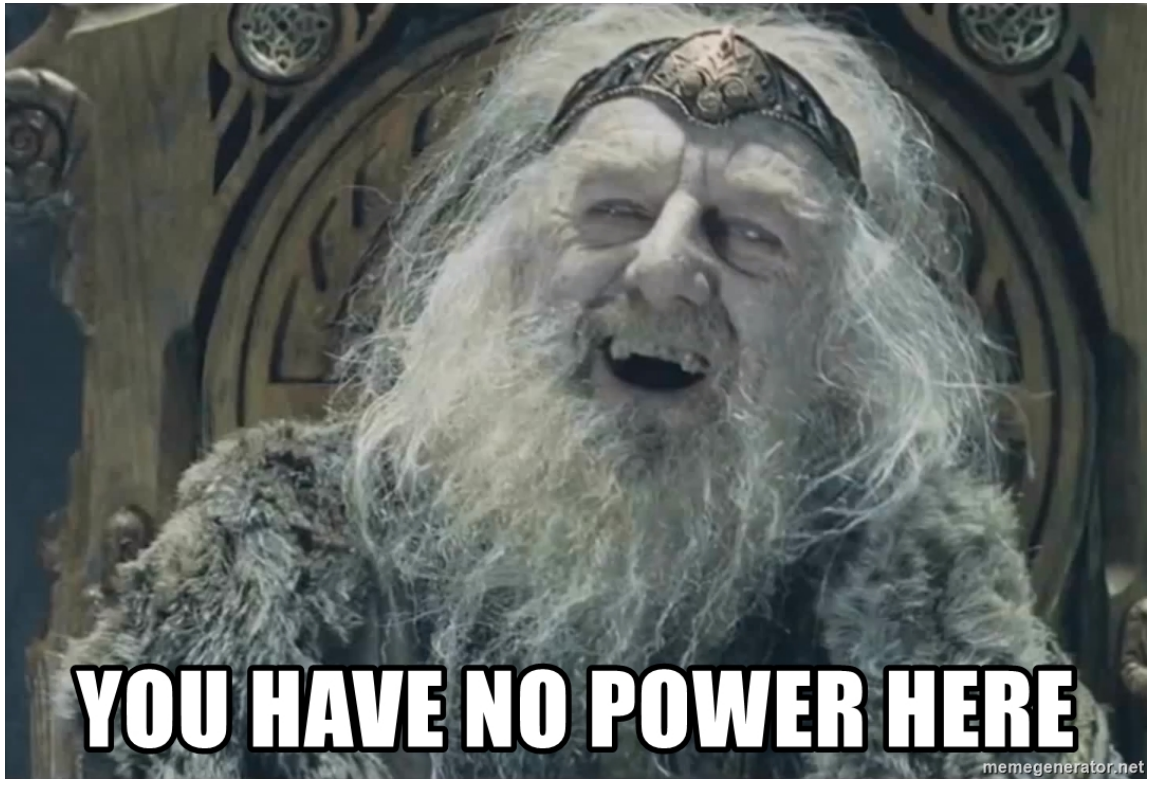
Replication reveal serious need for shrinkage.

The estimated effects size also have uncertainty: thus report confidence intervals.

Beware of small samples

Better to do a large replication than multiple small studies.

Otherwise, you risk being in this situation:



Observed (post-hoc) power

Sometimes, the estimated values of the effect size, etc. are used as plug-in.

The (estimated) effect size in studies are noisy!

The post-hoc power estimate is also noisy and typically overoptimistic.

Statistical fallacy

Because we reject a null doesn't mean the alternative is true.

When is this relevant? If the observed difference seem important (large), but there isn't enough evidence (too low signal-to-noise).