Power and model assumptions

Session 5

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

Outline

Power

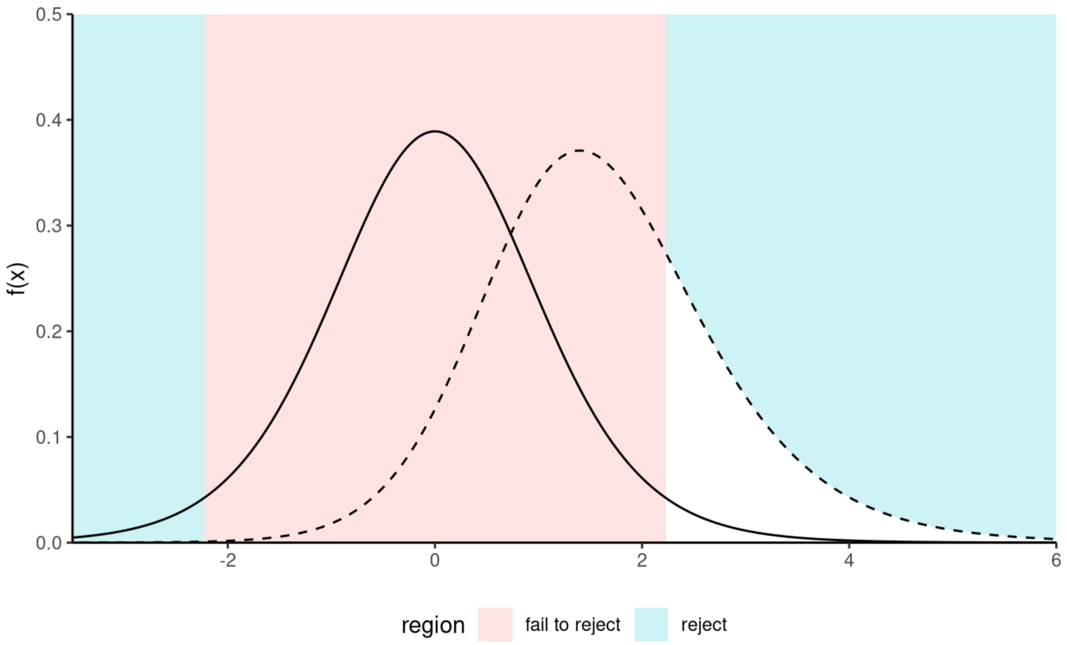
Effect and sample sizes

Model assumptions

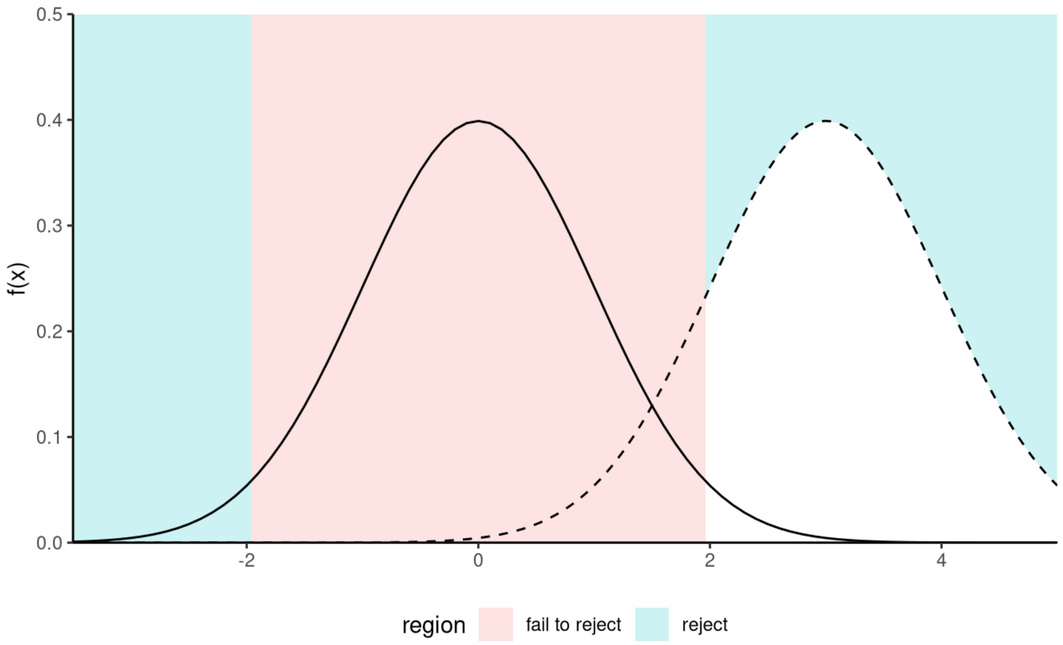
Power

I cried power!

The null alternative corresponds to a single value: $\mu_1 = \cdots = \mu_K$, 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 study replication: 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

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

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?

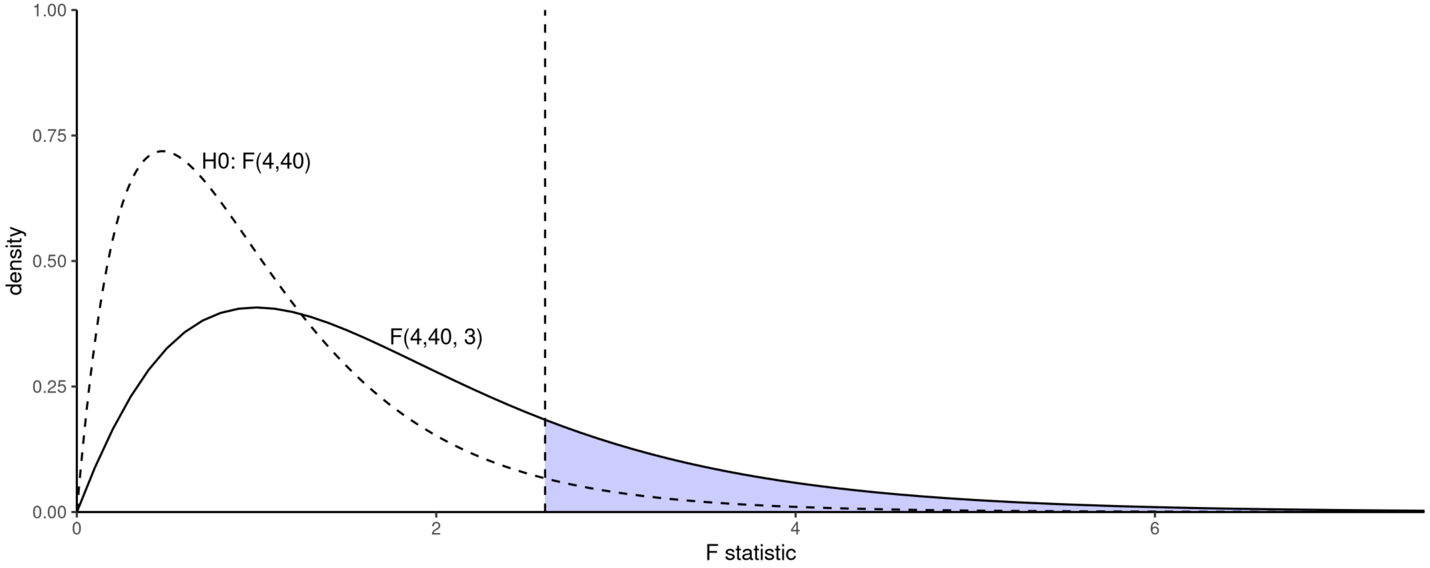
What is the effect of an increase of the

- 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 Δ, 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)
```

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?

How do we compute the power

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

Then,

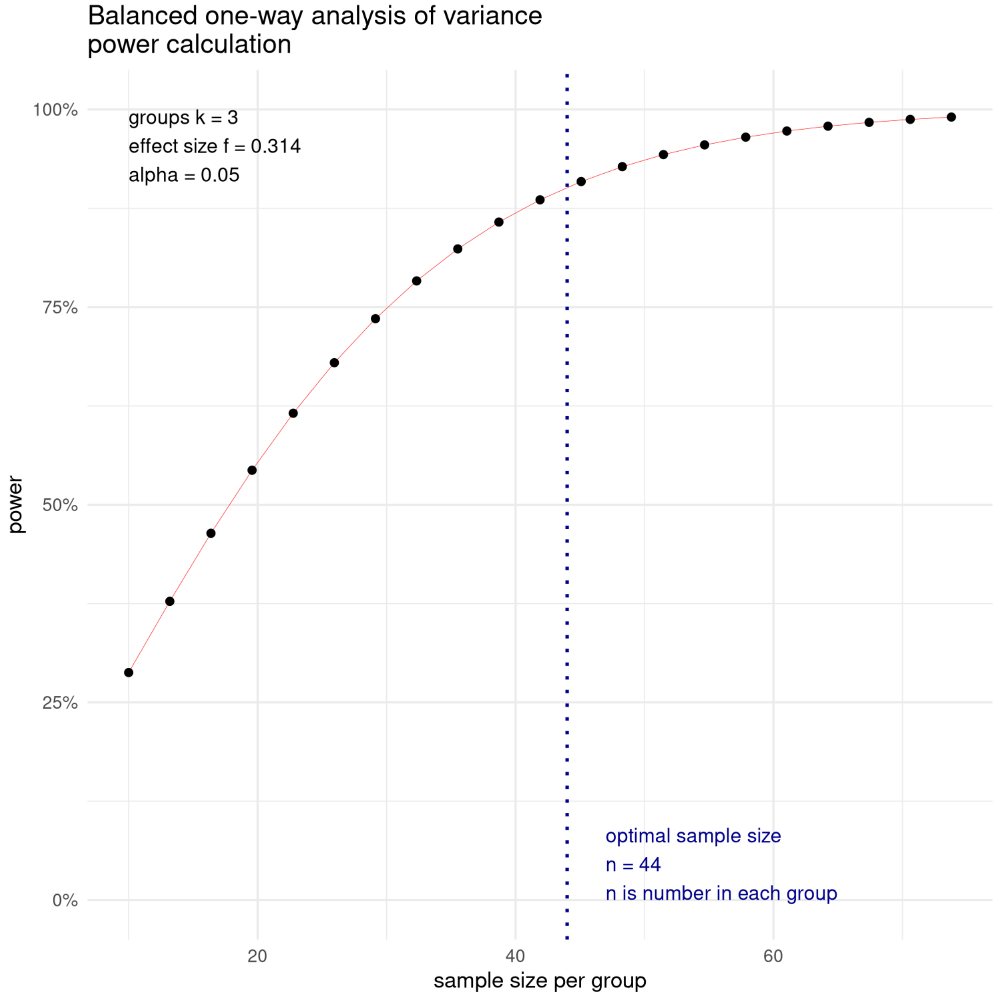
$$\Delta = \frac{n}{k\sigma^2} \sum_{j=1}^{k} (\mu_j - \mu)^2.$$

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

Power curves

```r
library(pwr)
power_curve <-
 pwr.anova.test(
  f = sqrt(0.09/(1-0.09)),
  k = 3,
  power = 0.9,
  sig.level = 0.05)
plot(power_curve)
```

Convert η^2 to Cohen's f (the effect size reported in `pwr`) via $f^2 = \eta^2/(1-\eta^2)$. Explained later in the slides.



Power calculations with unknowns

Need to specify the following parameters:

Level α (usually 5%)

adjust if you plan to correct for multiple testing!

Variance σ^2

run a pilot study or look up plausible values from the literature

Mean difference size $\mu_i - \mu$

Estimating mean differences

If you care for a difference of D units between groups, the smallest value of Δ (**worst case analysis**) is when

- one group has mean difference $\mu_1 - \mu = -D/2$,
- another has $\mu_2 - \mu = D/2$ and
- the rest with zero difference.

With a balanced design, we thus get

$$\Delta = \frac{n}{k\sigma^2}\sum_{j=1}^{k}(\mu_j - \mu)^2 = \frac{n}{k\sigma^2}\left(\frac{D^2}{4} + \frac{D^2}{4}\right) = \frac{nD^2}{2k\sigma^2}.$$

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).

Effect size

Measures of effects

Multiple alternatives

- standardized differences/measures
- percentage of variability explained (ratio of sum of squares or coefficients of determination).

They are all interrelated, but different people have different preferences.

Popularized in the handbook

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

Cohen's *d*

Popular standardized measure of effect (dimensionless=no units):

e.g., for a two-sample t-test between groups (comparing two means):

$$d = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\widehat{\sigma}}$$

where $\widehat{\sigma}$ is the pooled variance. 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 ($d = 0.8$) effect: these values change from one test to the next (above is for the two-sample t-test).

Eta-squared

Can break down the variability for sum of squares as 'total = within + between'. Define the percentage of variability explained by effect.

$$\eta^2 = \frac{\mathsf{SS}_{\text{effect}}}{\mathsf{SS}_{\text{total}}},$$

Problem: this depends on the variability of the measurement. Often, in more complex design, you see instead the following reported

$$\eta^2 = \frac{\mathsf{SS}_{\text{effect}}}{\mathsf{SS}_{\text{error}} + \mathsf{SS}_{\text{effect}}}.$$

For one-way ANOVA (no repeated measurements), the two are equivalent.

Eta-squared

For the balanced one-way ANOVA with exhaustive factors=all possible choices in the population, write

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

For the replication example,

$$\eta^2 = \frac{4.05 \times 2}{4.05 \times 2 + 84} = 0.088 \approx 0.09$$

Omega square

Yet another measure of effect size.

For one-way ANOVA, a transformation of the F-statistic gives

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

Partial versions for more complex designs, etc.

Uncertainty matters!

Every effect size is random (because inputs are)

Report confidence intervals with estimates.

Model assumptions