EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

February 27, 2020

Today

- "Standardized" measures of effect size
  - Cohen's d (pairwise differences, t-tests)
  - η^2 and partial η^2
  - (partial) ω^2
- Statistical power, a closer look
- Case study: Durante et al. 2013

Measures of effect size

- Always important to report effect sizes for any comparison, statistical test, or model. For example, *raw effect sizes*:
  - Observed difference in two sample means (t-test)
  - Two sample variances or standard deviations (F-test)
  - Sums of squares (main effects, interactions) in ANOVA
  - Regression coefficients in regressions
- Also need to report an estimate of *sample variability* (e.g. standard error, confidence interval, residual sum of squares)
- Cannot properly assess the meaning of an experiment/study without *at least* three things:
  - Observed effect size
  - Estimate of variability
  - Sample size(s)

"Standardized" measures of effect size

- Many statistics have been developed to try to communicate these three pieces of information in a single number.
- Unfortunately, this has the effect of obscuring easily understandable statistics (e.g. means, variances, counts) into obtuse derivative quantities (e.g. η^2, ω^2).
- Worst of all, since these derivative quantities are not immediately interpretable, people have developed rules of thumb for interpretation that now take the place of critical thought.
- Extremely common in social science literature (some health and natural sciences as well)

Cohen's d

- Cohen's d is the ordinary *standardized effect size* for the average difference between two groups:

$$d = \frac{\overline{X_1} - \overline{X_2}}{s},$$

  where s is an estimate of the overall standard deviation of the two groups.

- This is not an inherently bad statistic; if scales are different, standardizing can aide interpretation.

- However, Cohen's rule of thumb has become virtual gospel among applied practitioners. He advises:
  - $d \approx 0.2$ means small effect size
  - $d \approx 0.5$ means medium effect size
  - $d \approx 0.8$ means large effect size

Cohen's *d*

- **NEVER interpret your data this way,** at least, not without thinking hard if the interpretation is appropriate.
- First of all, it communicates nothing about *sample size*
- Secondly, the "small, medium, large" advice of Cohen only makes sense when *all your data are normally distributed*. Even mild deviations from normality can destroy these rules of thumb.
- Cohen's *d* is commonly reported with *t*-tests and post hoc pairwise comparisons from an ANOVA
- **Note:** Jamovi (and some other software) refer to Cohen's *d* as simply "effect size" - **be careful with the terminology**: it is only accurate to talk about a "raw effect size" (e.g. mean difference) and a "standardized effect size" (e.g. Cohen's *d*).

Eta-squared, η^2

- η^2 is another measure of "effect size": measures how much variation is explained by one factor (or one interaction) in an ANOVA:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

- Again, this is not an inherently bad statistic; we have been informally calculating it every time we look at an ANOVA table.
- However, the "proportion of total variance explained" interpretation only holds when:
  - group sizes are all equal (i.e. balanced ANOVAs)
  - there are no repeated measures (will study these ANOVAs soon)

Eta-squared, η^2

- η^2 can be useful for heuristics, but it can also hide a lot of important info:
  - Again, it communicates nothing about *sample size*.
  - It can hide the fact that your data don't explain much variation at all (e.g. SS_{total} is small).
  - Again, the (intuitive) interpretation breaks down for non-normal data.
- η^2 is always a *biased* estimator of the true variance explained.
- **Note:** η^2 for ANOVAs is the direct analogue of R^2 for regression models.
- Again, there are ill-advised rules of thumb for interpretation ($0.01 \approx$ small, $0.06 \approx$ medium, $0.14 \approx$ large): **NEVER use these.**

Partial eta-squared, $\eta^2_{partial}$

- $\eta^2_{partial}$ is a measure of how much variation is explained by one factor (or one interaction) relative to the residual variation:

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

- This is a bit more obscure (i.e. less intuitively interpretable) of a statistic.
- This is no longer "proportion of total variance explained" in any sense.
- This is a comparison of effect variance to residual variance.
  - Works when group sizes are not all equal (i.e. unbalanced ANOVAs)
  - Works with repeated measures (will study these ANOVAs soon)

Partial eta-squared, $\eta^2_{partial}$

- $\eta^2_{partial}$ hides and obscures a lot of important info:
  - Again, it communicates nothing about *sample size*.
  - Again, it can hide the fact that your data don't explain much variation at all.
  - Again, the (intuitive) interpretation breaks down for non-normal data.
  - It will *automatically increase* as you add more terms to your ANOVA model, since the leftover variation, SS_{error}, will automatically go down.
- $\eta^2_{partial}$ is again always a *biased* estimator of the true variance explained.
- **Note:** $\eta^2_{partial}$ for ANOVAs is analogous to $R^2_{partial}$ for regression models.
- Again, there are ill-advised rules of thumb for interpretation (0.01 $\approx$ small, 0.06 $\approx$ medium, 0.14 $\approx$ large): **NEVER use these.**

Omega-squared, ω^2

- (partial) ω^2 is a measure of how much variation is explained by one factor (or one interaction) relative to the total and residual variation:

$$\omega^2 = \frac{SS_{effect} - df_{effect} \cdot MS_{error}}{SS_{total} + MS_{error}}$$

- This is a *lot* more obscure of a statistic.
- It tries to again mimic the "variance explained by the effect of interest" paradigm.
- This is a comparison of effect variation to total and residual variation.

Omega-squared, ω^2

- ω^2 hides and obscures a lot of important info:
  - Again, it communicates nothing about *sample size*.
  - Again, it can hide the fact that your data don't explain much variation at all.
  - Again, the (intuitive) interpretation breaks down for non-normal data.
- ω^2 is again always a *biased* estimator of the true variance explained, although not as badly biased as η^2 or $\eta^2_{partial}$.
- **Note:** ω^2 for ANOVAs is analogous to $R^2_{adjusted}$ for regression models.
- Again, there are ill-advised rules of thumb for interpretation ($0.01 \approx$ small, $0.06 \approx$ medium, $0.14 \approx$ large): **NEVER use these.**

Obscure effect size measures for our toy example

Recall two-way ANOVA model, with interaction, for Anxiety vs. Education and Sex:

$$Y_{anx} = \mu + \tau_{edu} + \tau_{sex} + \tau_{edu \times sex} + \varepsilon$$

ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	$\eta^2 p$	ω^2
Education	10.294	2	5.147	63.187	<.001	0.644	0.940	0.631
Sex	0.011	1	0.011	0.140	0.718	0.001	0.017	–0.004
Education ✻ Sex	5.023	2	2.511	30.830	<.001	0.314	0.885	0.303
Residuals	0.652	8	0.081					

- Note that all these different "effect size" measures give no greater insight than simply reporting the original SSs or MSs (effects and residual); in fact, they give the same info as the F-statistics.
- In fact, they simply replace easily interpretable quantities (sample variances) by obscure decimals.
- **Advice: report these statistics only if required by a journal.**

Statistical power

- The concept of *statistical power* is crucial for both designing a study and for interpreting a study that has already been conducted.
- *Power* is (informally) defined as the ability to detect non-zero effects (true positives)
- The *power*, or *sensitivity*, of a test is defined as

$$\Pr(p - value < \alpha \mid H_0 \text{ false}) = 1 - \beta,$$

where α is the *significance level* set by the researcher/journal and used to declare p-values "significant" or not under the traditional threshold approach.

- Good studies will strive to have $1 - \beta \geqslant 0.80$. Most studies will have much lower power.

Statistical power

	H_0 true	H_0 false
data inconsistent with H_0	Type I error *false positive*	Correct decision *true positive*
data consistent with H_0	Correct decision *true negative*	Type II error *false negative*

	Given H_0 true	Given H_0 false
Pr(data inconsistent with $H_0 \mid \cdots$)	α	$(1-\beta)$
Pr(data consistent with $H_0 \mid \cdots$)	$(1-\alpha)$	β

Statistical power

- Statistical power is a function of many things:
  - Sample size (increasing sample size automatically increases power)
  - Population variability (less variation means more power)
  - Overall distribution of random phenomenon of interest (average effects in clustered or multi-modal distributions can be difficult to detect)
  - Type I error rate, α (increasing α automatically increases power)
  - *True, unobserved effect size* (bigger effect sizes are easier to find)
  - Type of statistical test/procedure used (e.g. nonparametric or robust procedures can be more powerful when data are non-normal)
  - Measurement error (noisier measurements produce more variability, so lead to less power)

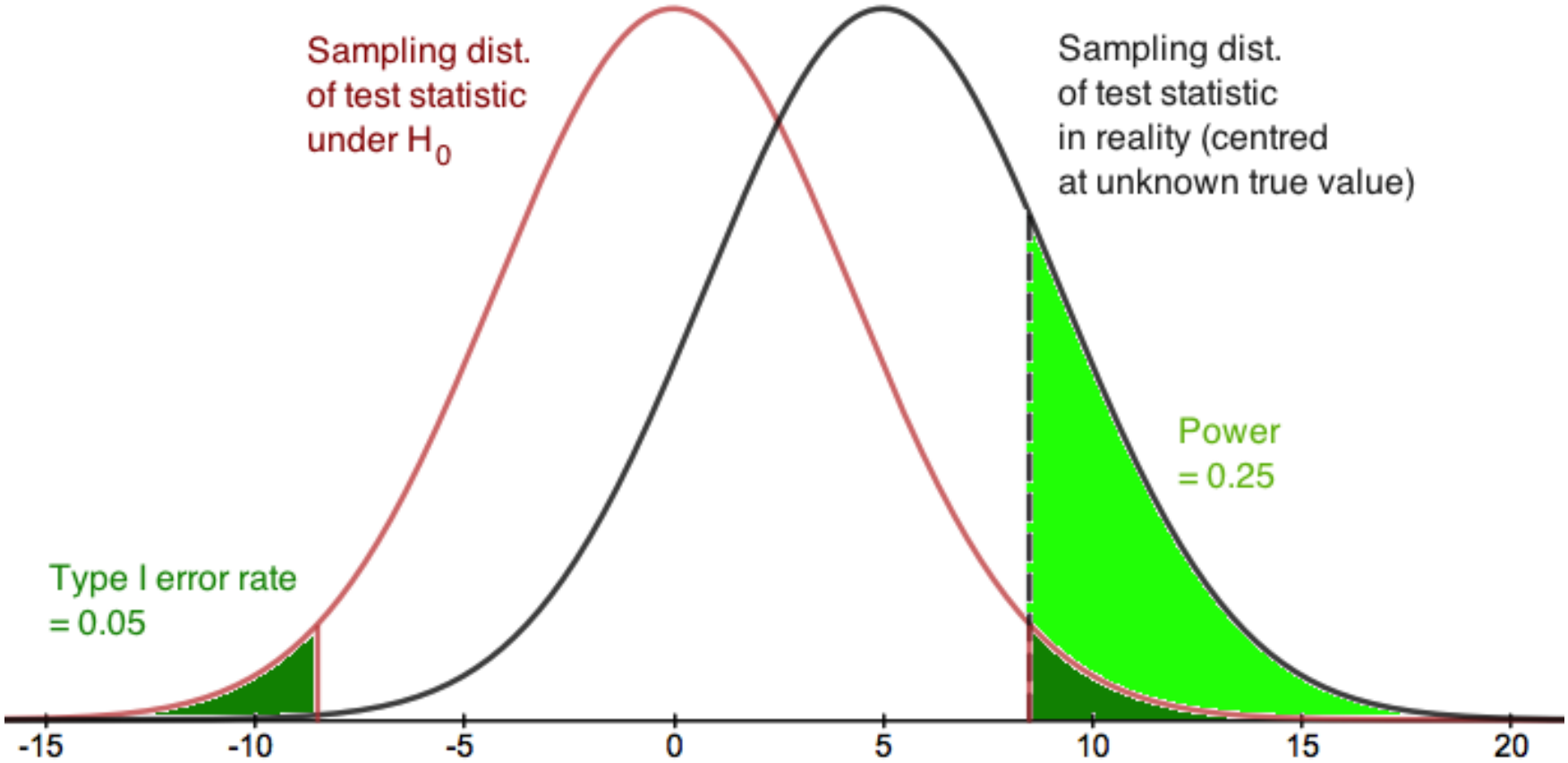
Statistical power

- When planning a study, power is considered to determine how large your *sample size* should be. This is called *power analysis* and generally proceeds as follows:
  - Identify the goal of the research study (e.g. testing if a new drug or intervention is more effective over current treatments)
  - Identify how you will measure the outcomes, effect size (e.g. mean difference between two treatment groups)
  - Use the previous literature to *reasonably estimate the variability* in your future study (e.g. similar drugs tested produced about a σ^2 variation in the response)
  - Decide on how you will analyze your outcomes (e.g. t-tests, ANOVAs, regression)
  - *Determine what effect size would be clinically important enough for you to care* (e.g. you want a new drug to be at least 20% more effective than current treatments)
  - Set your type I error rate α.
  - Set your desired power $1 - \beta$; i.e. your desired ability to detect the effect of clinical importance to you.

Statistical power

- Only after all this setup can we then estimate the necessary sample size to attain the desired power (more next time).
- This is a necessary step of virtually all medical research.
- This is often a necessary step to obtain funding for a proposed project. Why?
  - If you design a study that has a poor chance of detecting what you are trying to find, then why bother doing the study at all?
  - If your study has low power, but you end up finding a significant non-zero effect anyway, *it is likely that you are making a type I error*.
  - Moreover, if your study has low power but you end up finding a significant non-zero effect anyway, your *effect estimates are likely massively overinflated* (Type S and Type M errors).

Statistical power

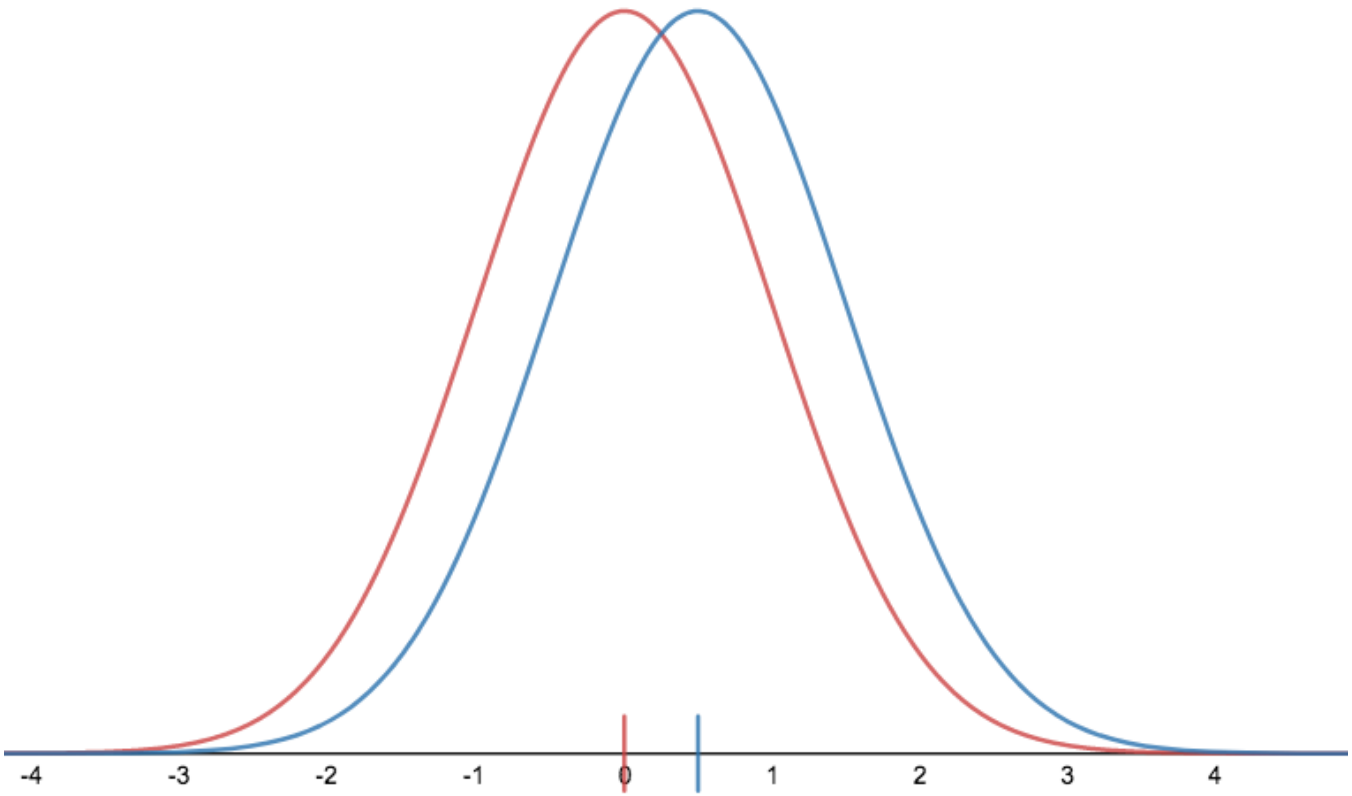


- Should *always* have this picture in mind when thinking about power.

Examples of study situations with different powers

Note: "small" and "large" are *relative* terms

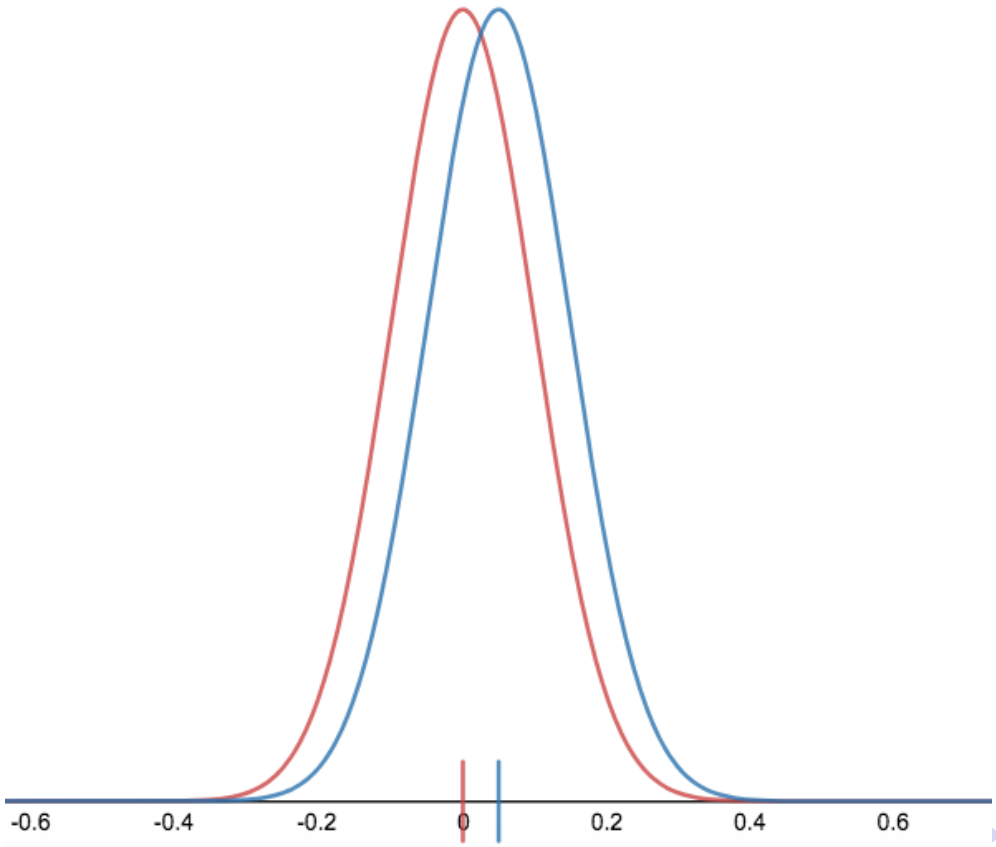
- Low power
- Small true effect size (0 vs. 0.5)
- Small sample size and/or large variance



Examples of study situations with different powers

Note: "small" and "large" are *relative* terms

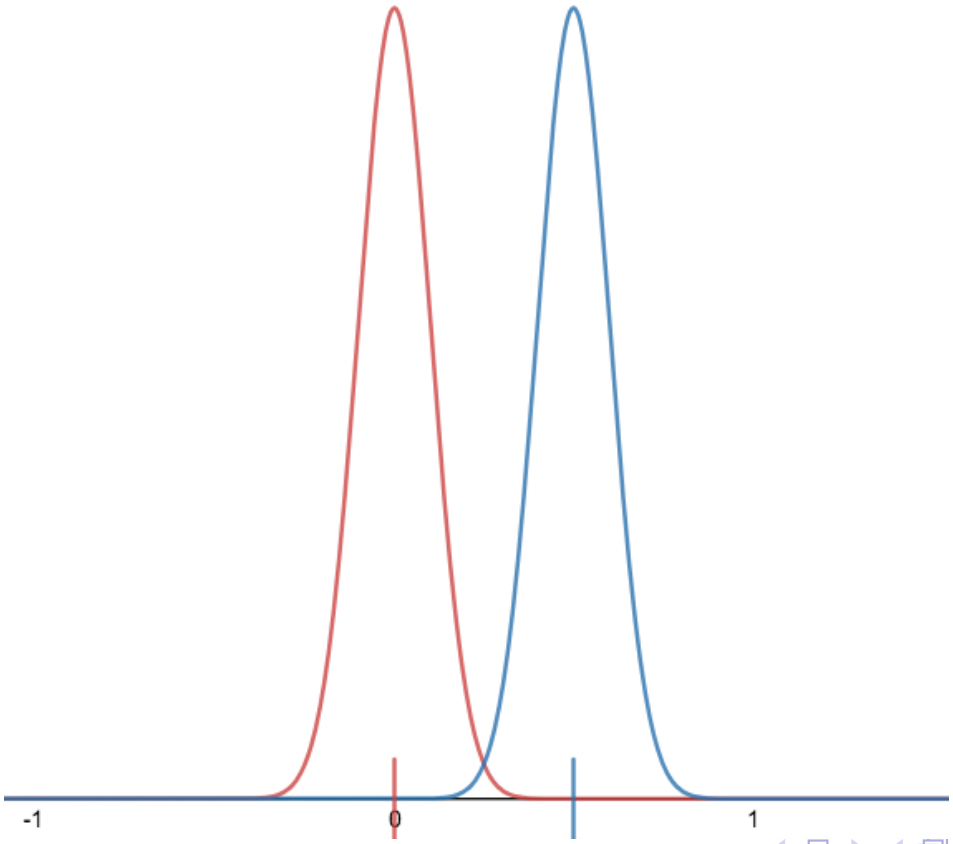
- Low power
- Small true effect size (0 vs. 0.05)
- Large sample size and/or small variance



Examples of study situations with different powers

Note: "small" and "large" are *relative* terms

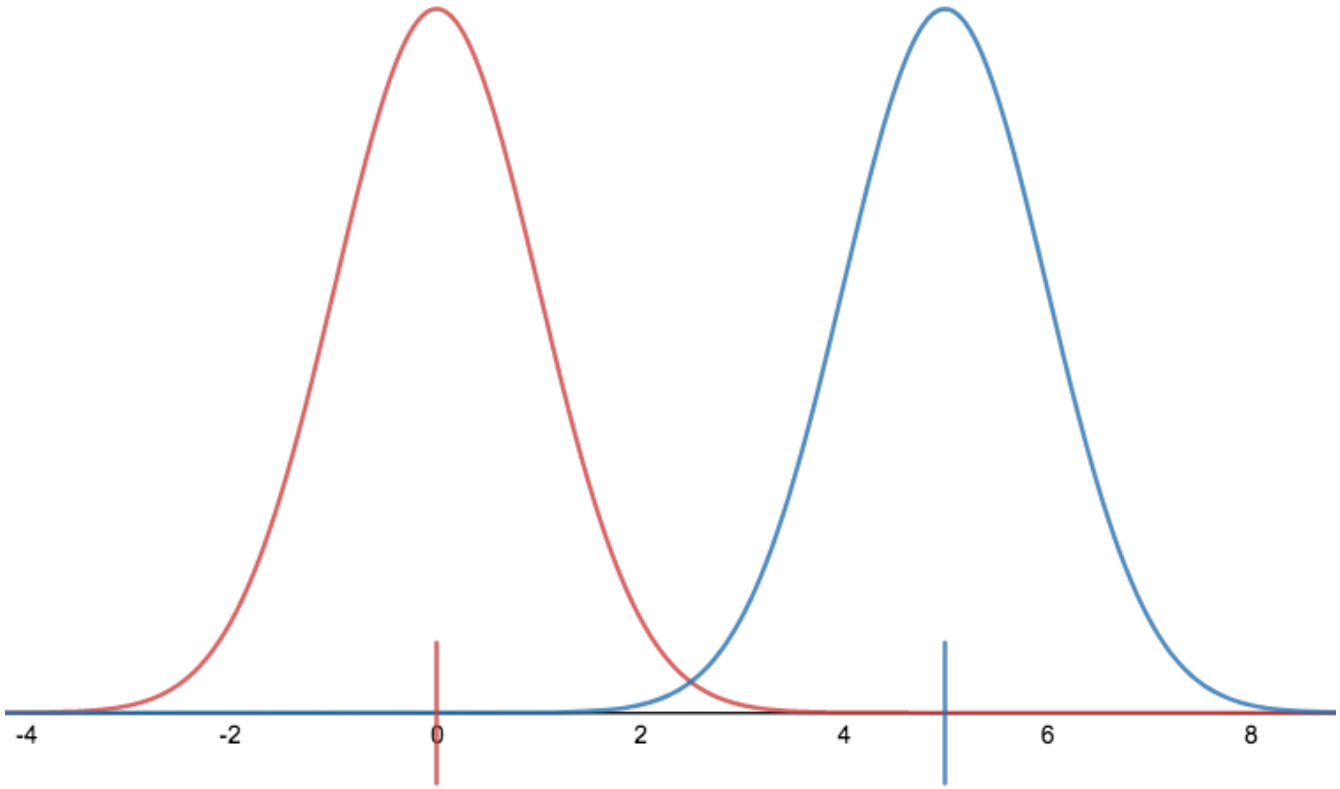
- High power
- Small true effect size (0 vs. 0.5)
- Large sample size and/or small variance



Examples of study situations with different powers

Note: "small" and "large" are *relative* terms

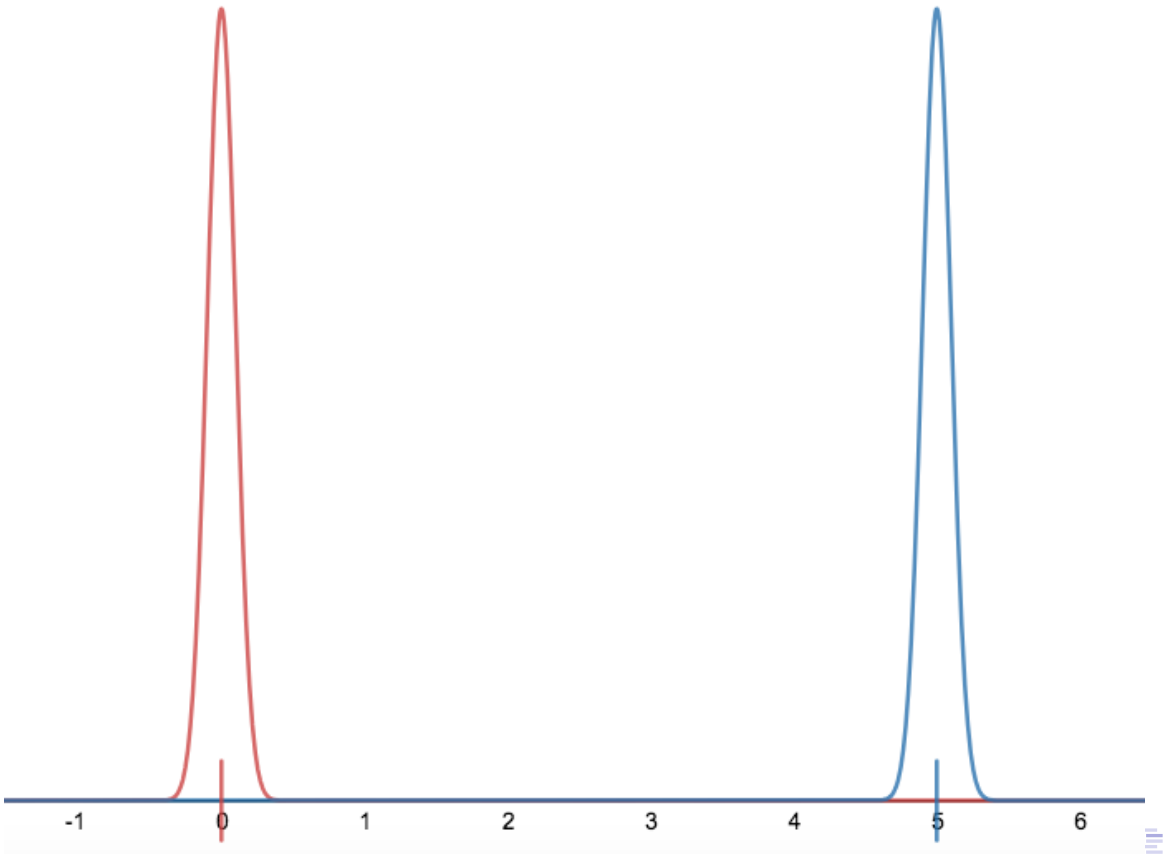
- High power
- Big true effect size (0 vs. 5)
- Small sample size and/or large variance



Examples of study situations with different powers

Note: "small" and "large" are *relative* terms

- Really high power (won't even require a statistical test of hypotheses)
- Big true effect size (0 vs. 5)
- Large sample size and/or small variance



How to calculate statistical power

- For simple scenarios, power can be calculated analytically (i.e. by hand). **But we rarely study simple scenarios.**
- *Lots* of software exists that claims to calculate power for you (e.g. SPSS, G*Power); but all of it relies on *the simple scenarios that rarely apply in practice*.
- In particular, software nearly always relies on an assumption of *perfectly normal data*; see Oscar Olvera Astivia's blog post.
- Practically, this means that sample size estimates can be grossly distorted (very, very bad!)
- Usually no software or analytical options available for complicated study designs.
- What to do?

How to calculate statistical power

- What to do? **Must simulate (i.e. perform a simulation study) to perform power analysis.**
- Simulation allows you to tailor a sample size estimate to the exact specifics of any study design.
- Simulation requires semi-decent programming capabilities.
- If you don't have these skills, *seek a statistician's help!*

Effects of low power on interpretation of analytical output

Low power can come from many different sources. In practice, the three most common are:

- Small sample sizes (overall, or within groups).
- Large variability (overall, or within groups, or due to noisy measurements).
- Small *true* effect sizes.

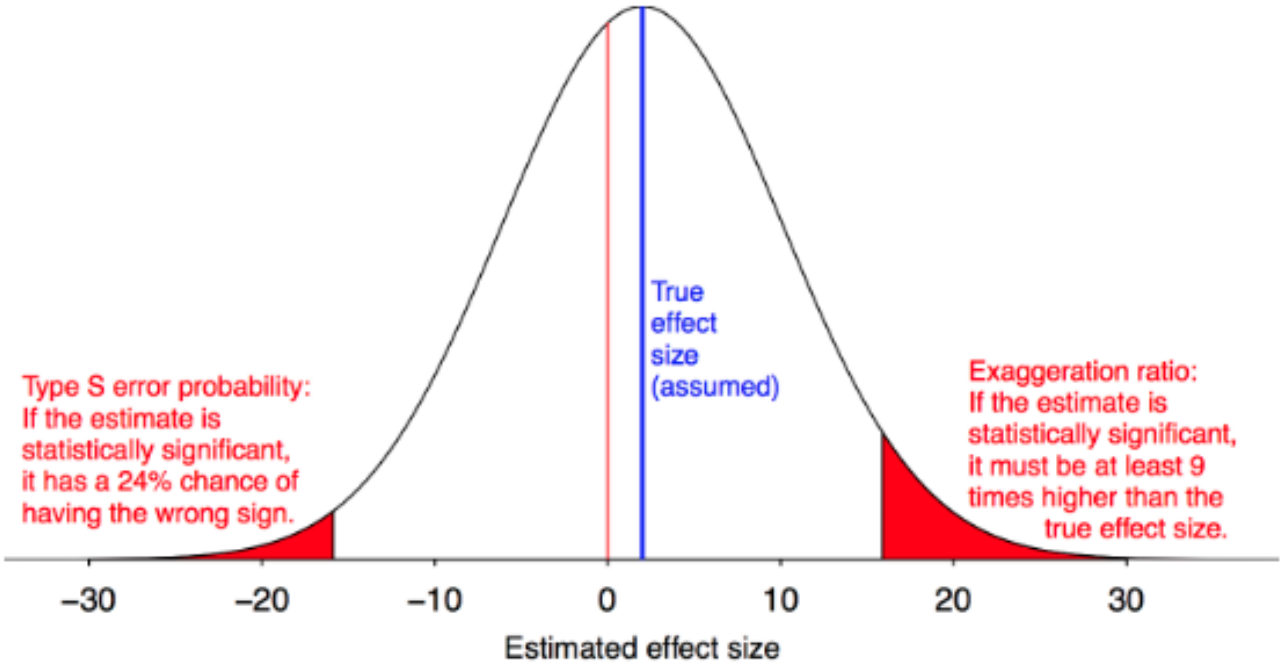
The first two sources are easy to see. The last (small true effect sizes) is difficult and subjective, but absolutely crucial.

Effects of low power on interpretation of analytical output

True effect sizes are *unobserved*, but crucial to interpretation:

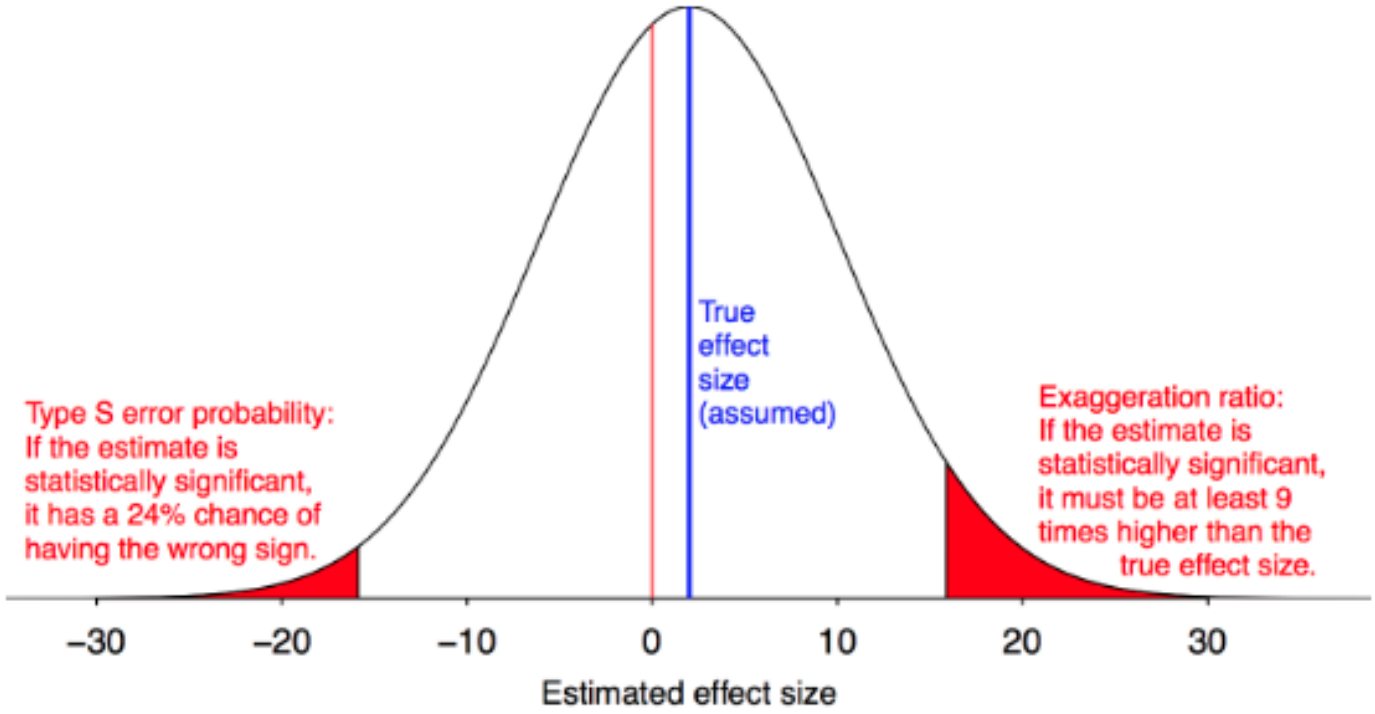
- We never actually know the *true* effect size (if we did, we wouldn't have to perform a study to estimate it).
- A plausible true effect size depends on the *prior believability of a particular alternative hypothesis*.
- In social science, many of our effects of interest will be small, *especially when compared to the effects of other variables of little or no interest*.
- *Evaluating the power of a study retrospectively requires an informed assessment of how plausible you would find certain effect sizes.*
- **Note:** some applied practitioners and software (e.g. SPSS) will talk about "retrospective power" or "post hoc power analysis"; they do *not* mean what we are talking about (usually, they mean gibberish).

Effects of low power on interpretation of analytical output



- This is a graphical representation of a t-test comparison of means.
- The *statistical power* here is 6%.
- In this example, true effect size (marked by blue line) is very small.
- Red regions represent values for "significant" test statistics (and so, p-values)

Effects of low power on interpretation of analytical output

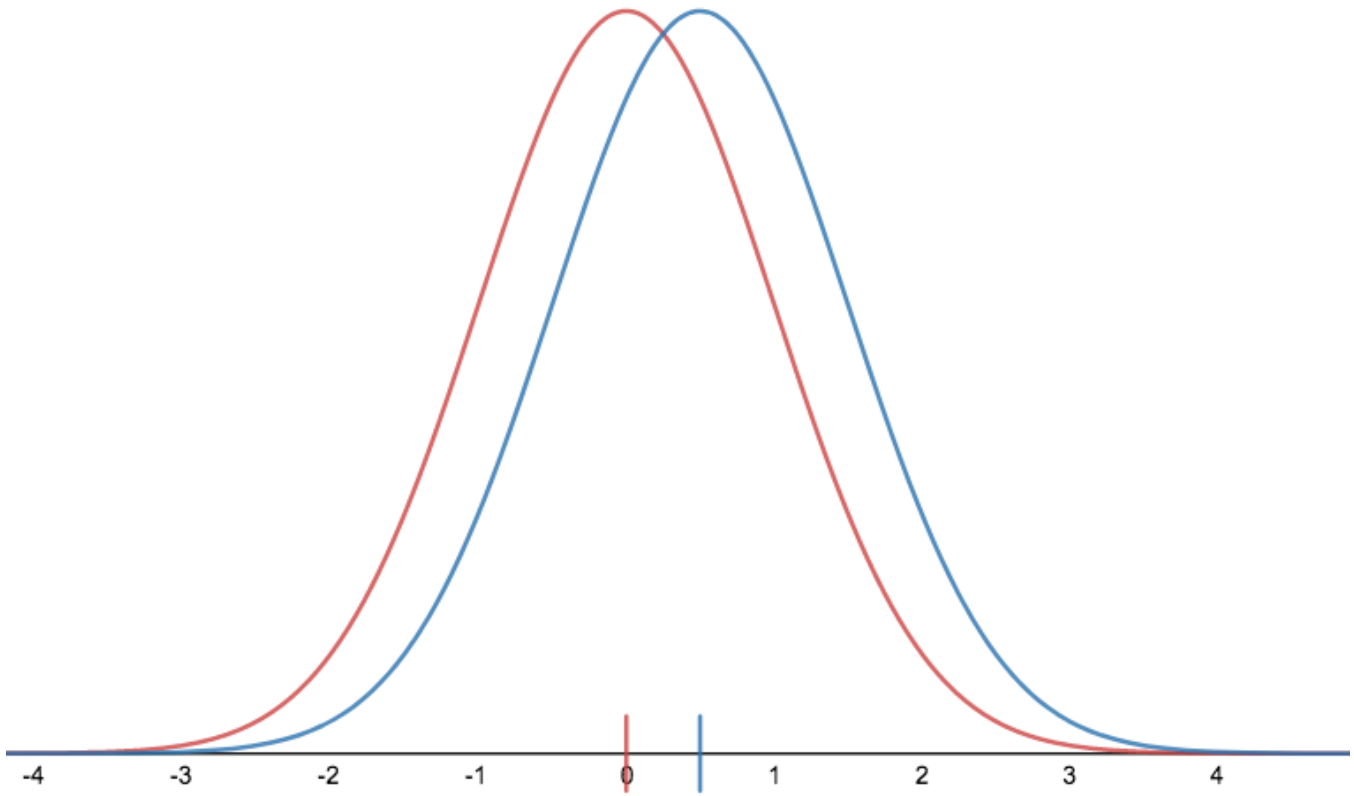


- But then finding a significant result would mean:
  - the estimated effect size is at least 9 times too big (Type M error)!
  - the estimated effect size has the wrong sign about 25% of the time (Type S error)! [See Gelman & Carlin (2014) for more info.]

Examples of study situations with different powers

Note: "small" and "large" are *relative* terms

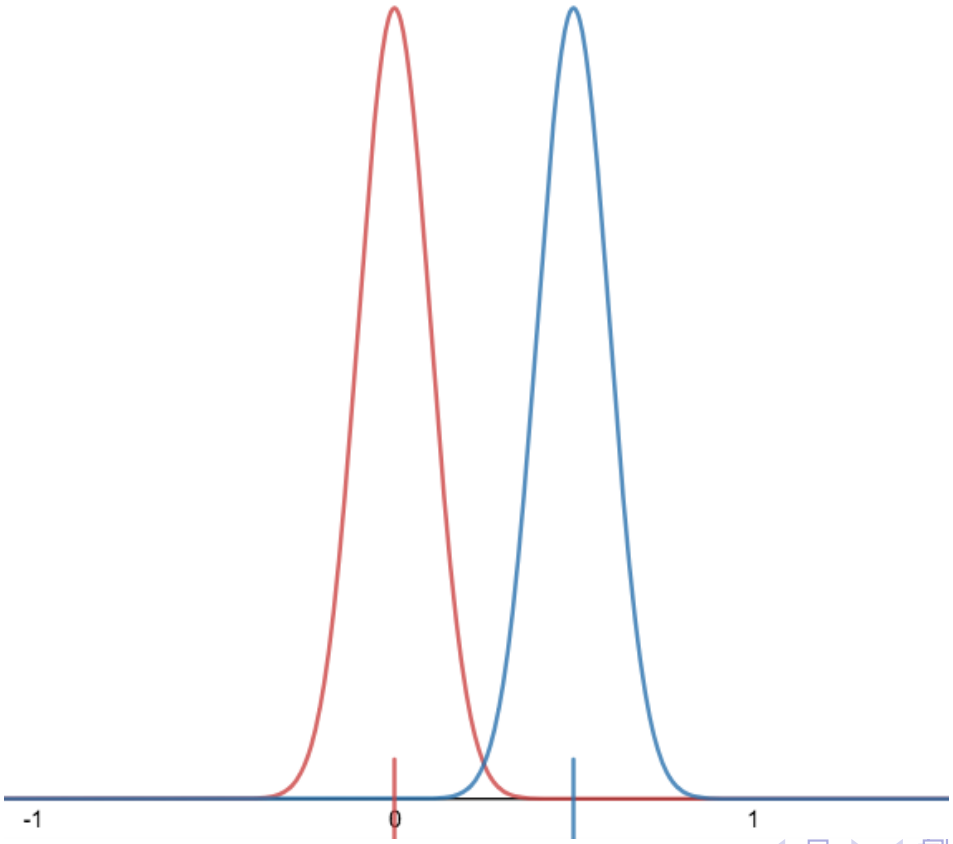
- Low power = bad estimates if significant
- Small true effect size (0 vs. 0.5)
- Small sample size and/or large variance



Examples of study situations with different powers

Note: "small" and "large" are *relative* terms

- High power = good estimates if significant
- Small true effect size (0 vs. 0.5)
- Large sample size and/or small variance



Effects of low power on interpretation of analytical output

In low-powered studies:

- Significant results are often meaningless.
- Significant results *will* yield estimates that are wildly inaccurate.
- Seemingly small things like measurement error, sampling variability, or minor experimental imperfections become magnified.
- Results are often entirely driven by statistical "noise".

Case study

- Case study: Durante et al. 2013