

BIO708 February 15th, 2024

effect sizes

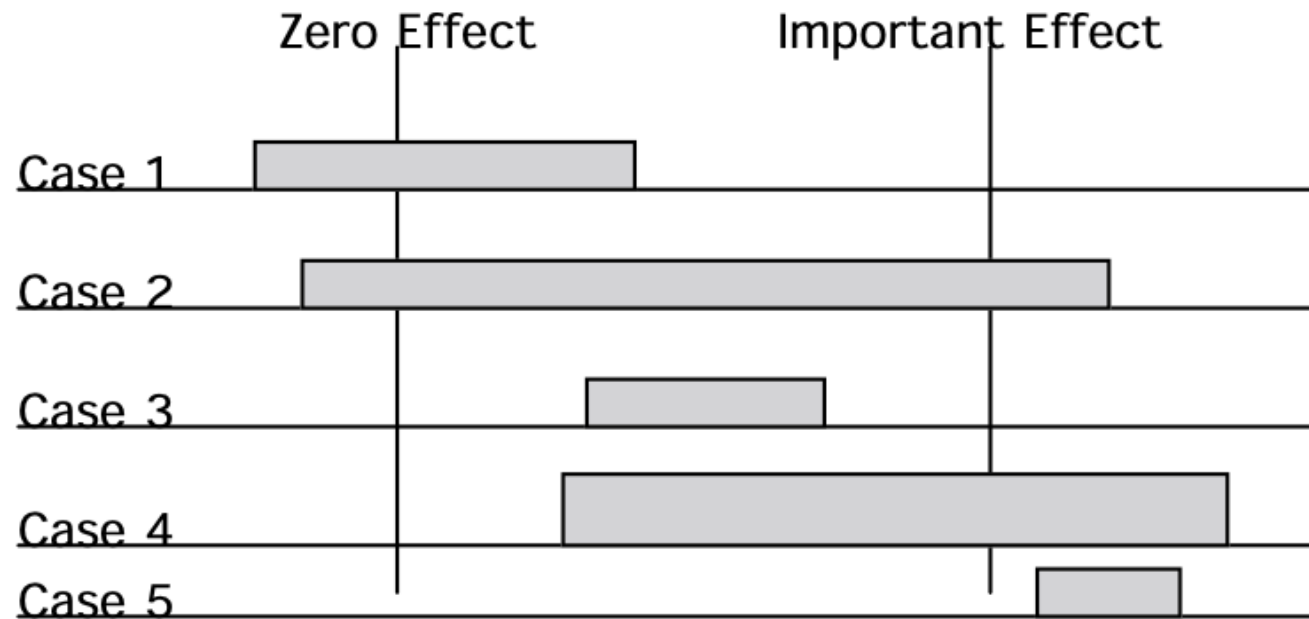


Figure 1. Confidence intervals for five different environmental scenarios.

Biological VS statistical significance and the importance of effect sizes

The importance of effect sizes, and why test statistics
are not effect sizes.

Goals for the day

Discuss the idea of effect sizes, and why (with Confidence Intervals) they are central to any statistical, and biological, inferences.

By the end of class

- You will be able to
 - Distinguish between statistical significance testing and effect sizes.
 - Distinguish and determine (for your own work) between non-standardized measures of effect sizes, and standardized measures of effect (standardized based on standard deviation, or mean

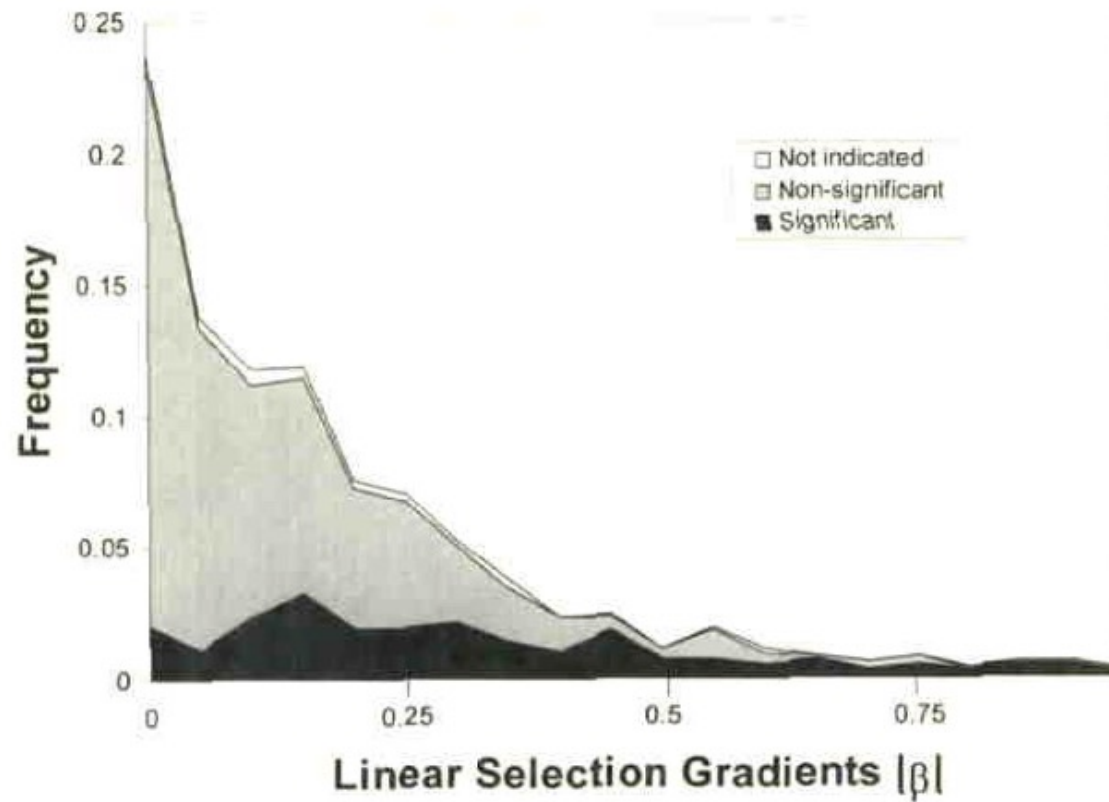
Readings for effect sizes

- A biologically focused paper is [Nakagawa and Cuthill 2007](#). This has all relevant bits, including sample size corrected d , and how to convert between d and r).
 - *I personally like this one a great deal.*
- [Jané et al. 2023. Guide to effect sizes and confidence intervals](#).
 - A useful and practical guide to effect sizes, also with links to R packages and example scripts. Best to read and use after Nakagawa and Cuthill.
- Cumming book has a good introduction to effect sizes. Back end of chapter 2 on effect sizes, and chapter 11 on Cohen's d and related measures.

R Libraries (probably lots more)

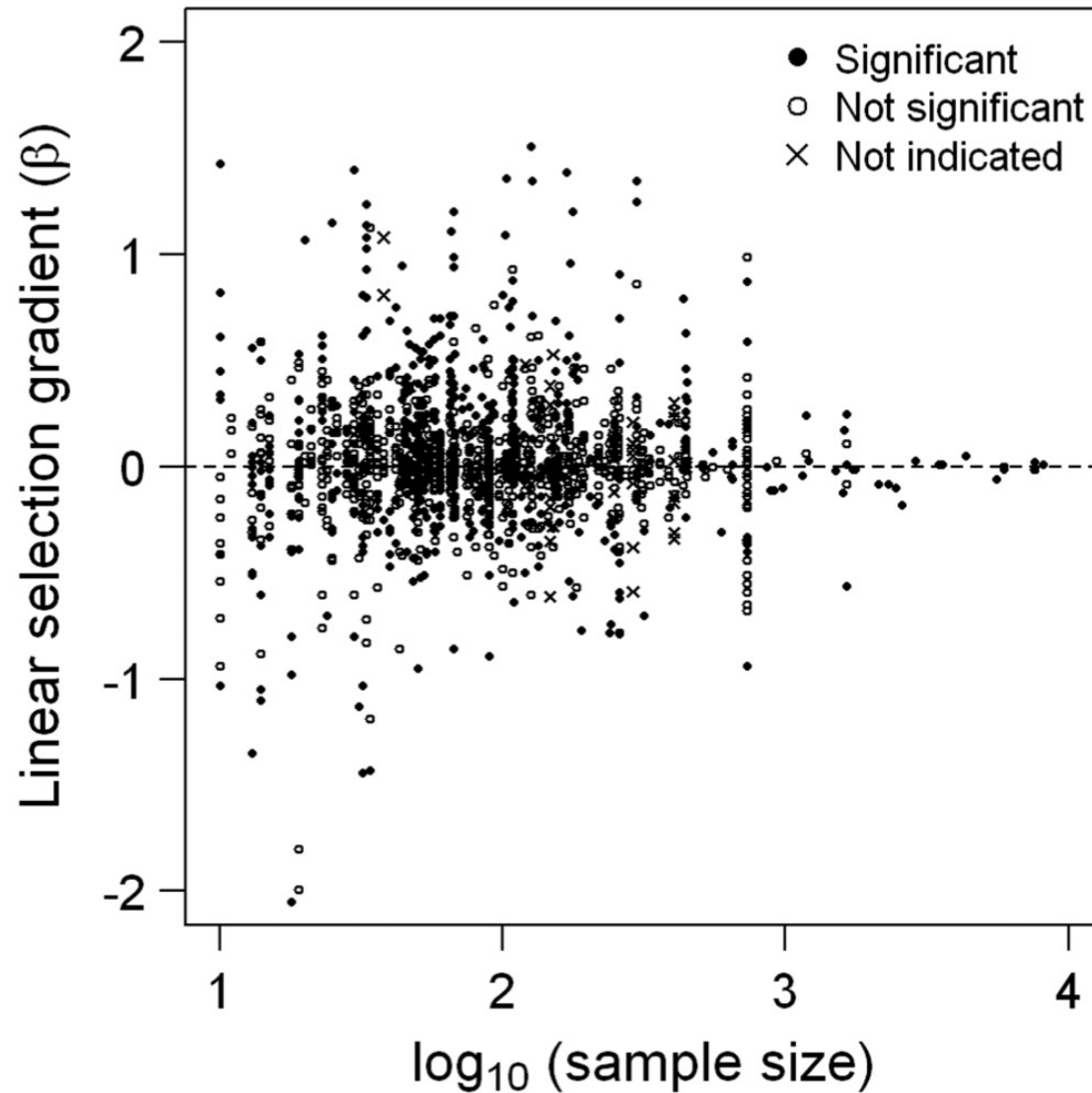
- effectsize (<https://easystats.github.io/effectsize/reference/index.html>)
- effsize (<https://cran.r-project.org/web/packages/effsize/effsize.pdf>)
- esviz (<https://github.com/datalorax/esviz>)
- dabestr
- emmeans, which we will use a lot when we introduce linear models can also compute many common effect sizes, and their confidence limits.

Motivation: Is directional selection strong in nature?



Kingsolver et al 2001

Motivating example: The strength of natural selection in the wild



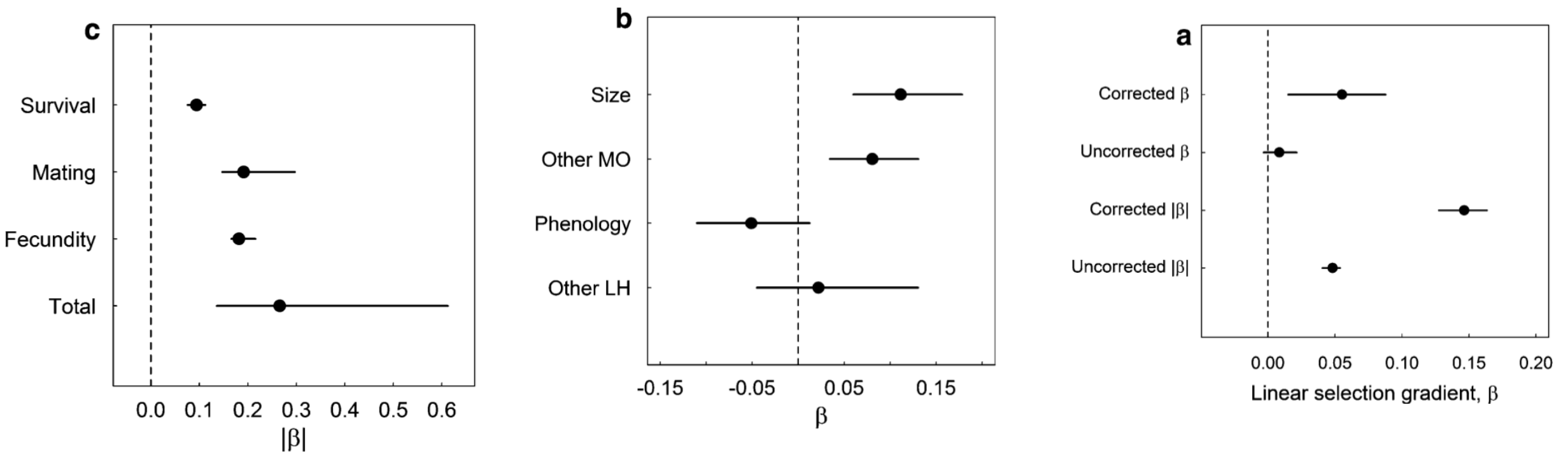
These are measures of the strength of selection from hundreds of studies.

- How is it that estimates from all of these studies can all be compared?
- How do we determine if the strength of selection is strong?

In the field of phenotypic evolution, early (largely theoretical) researchers (Lande 1976, and Lande & Arnold 1983), followed by several empirical researchers (Conner, Janzen, Brodie, Schluter... many more) decided that the continuous predictors (targets of selection) should be z-transformed (centered and scaled by standard deviation) and response variables (fitness proxies) should be scaled by the mean.

Most researchers followed suit, allowing everyone to compare the effects of selection in a standardized way

This allowed meaningful comparisons across many studies and really helped to move the field forward in terms of understanding the strength of selection

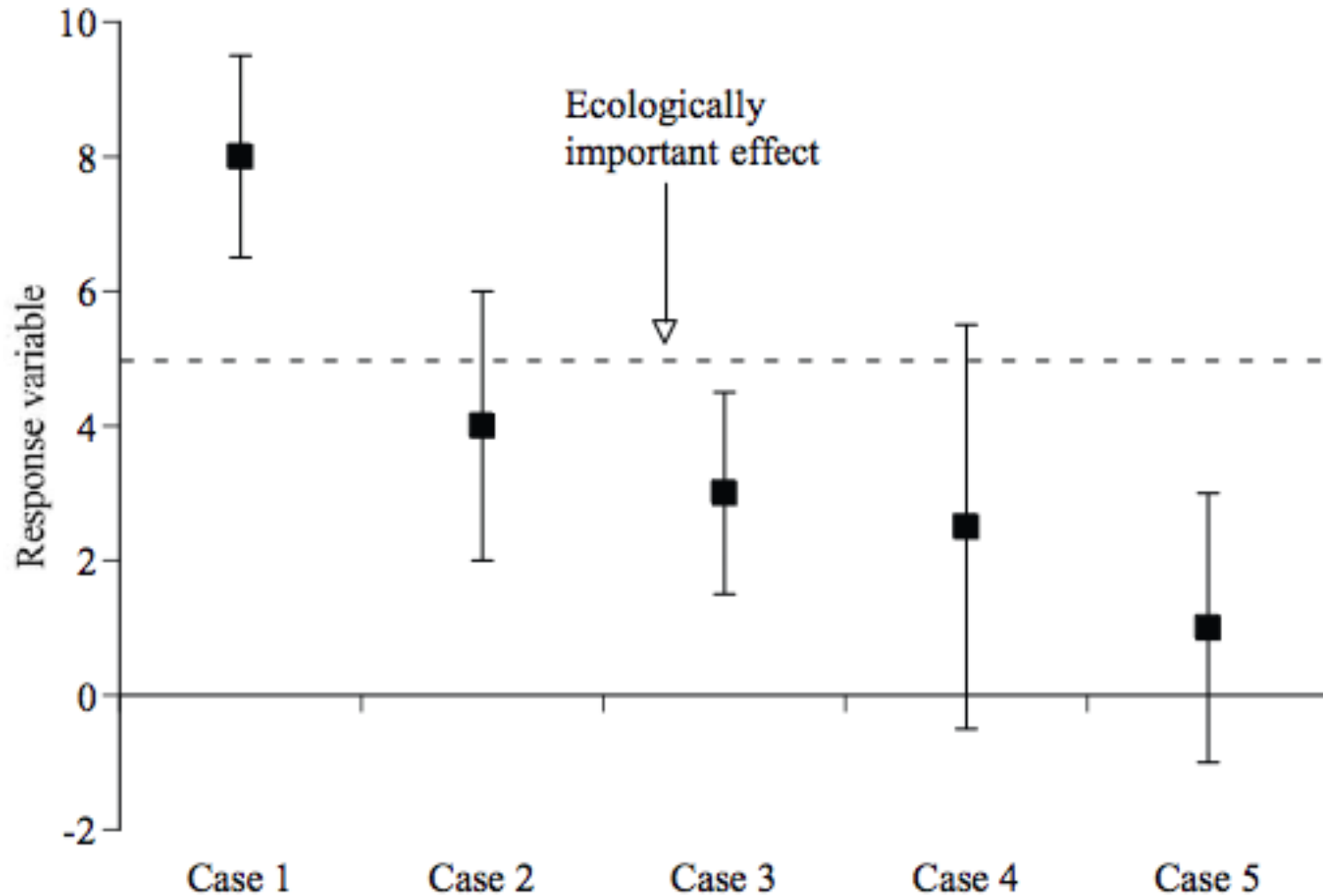


From 1341 estimates of selection gradients (β)

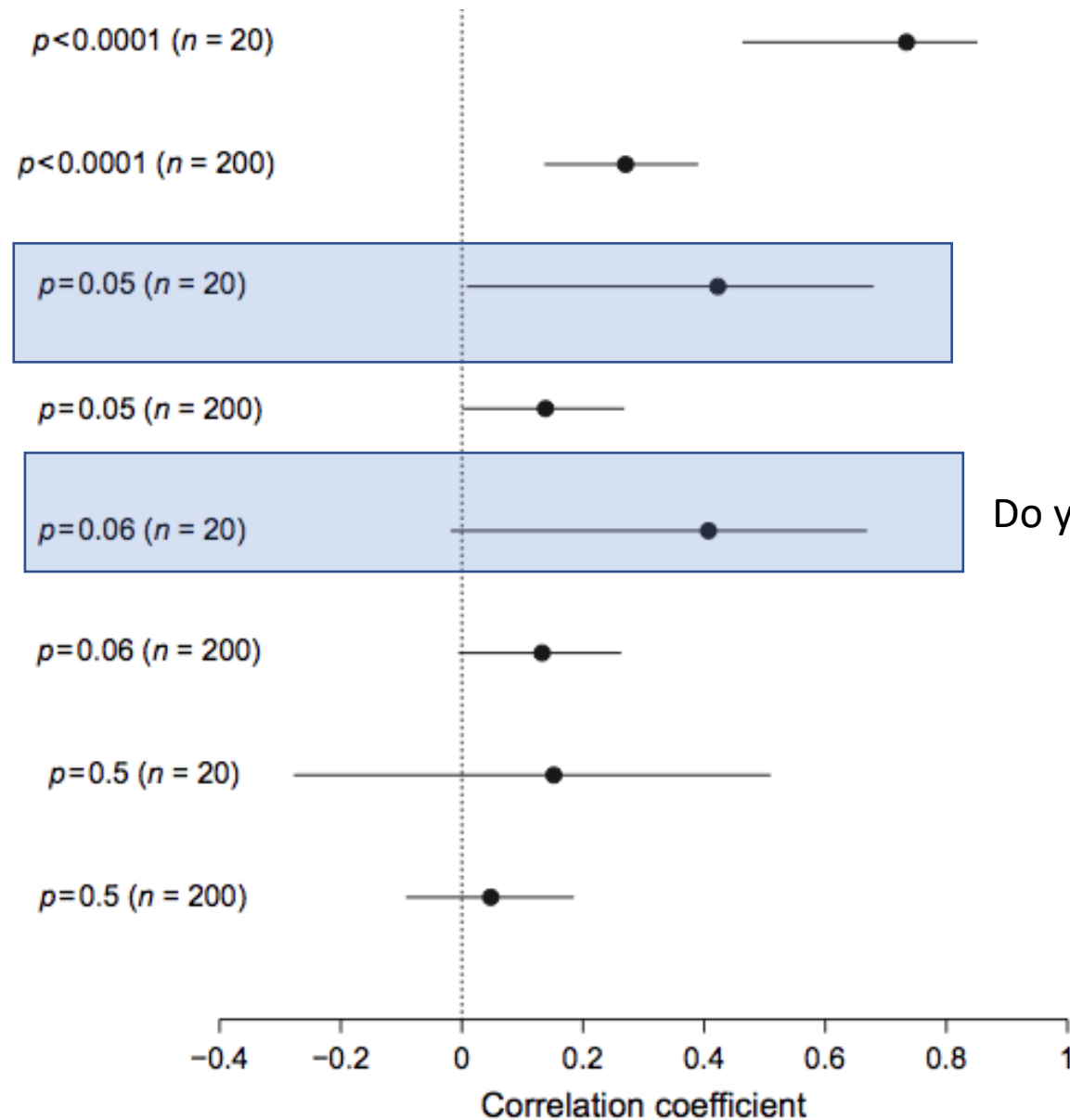
Kingsolver et al 2011

Why focusing on effect sizes (and CIs) can be so helpful

It may be significant, but is it important?



NHST and the problem of shrinking estimates to 0



Do you really want to treat this estimate as 0?

The Big Picture

- The coefficients (estimated parameters) from our models are not simply estimates to be examined along with p-values, but are the ***most important aspect of the model*** with respect to your ability to assess the importance of particular variables.
- Frame your questions around “What is the effect?”,
 - **Not** around the question “Is there an effect?”

Salient points of the material

- There are several classes of effect sizes (unstandardized, scaled by pooled sd , scaled by mean, variance accounted for, odds ratios).
- Deciding which ones to use (*a priori*) depends on the questions at hand, and what you plan to compare your results to (and practical importance).
- This can take a considerable amount of thought. But it will help you so much in being able to interpret your results.

t-test review (two sample t-test)

$$t = \frac{\bar{x}_F - \bar{x}_M}{s_p \sqrt{\frac{1}{n_F} + \frac{1}{n_M}}}$$

$$s_p = \sqrt{\frac{(n_F - 1)s_F^2 + (n_M - 1)s_M^2}{n_F + n_M - 2}}$$

Pooled standard deviation

t-test review (two sample t-test): numerator is the difference

- For two groups (males and females), we want to estimate the mean difference between them.
- But what else do we need to account for?

$$diff = \bar{x}_F - \bar{x}_M$$

How about just the difference between
the means?

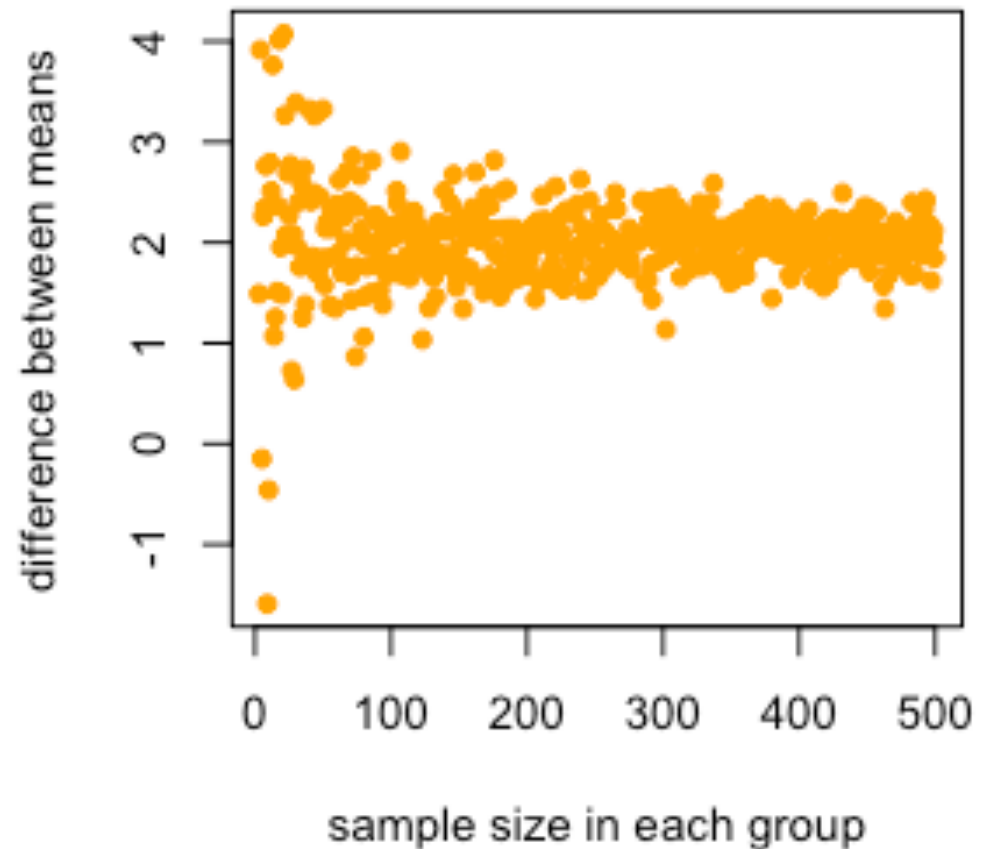
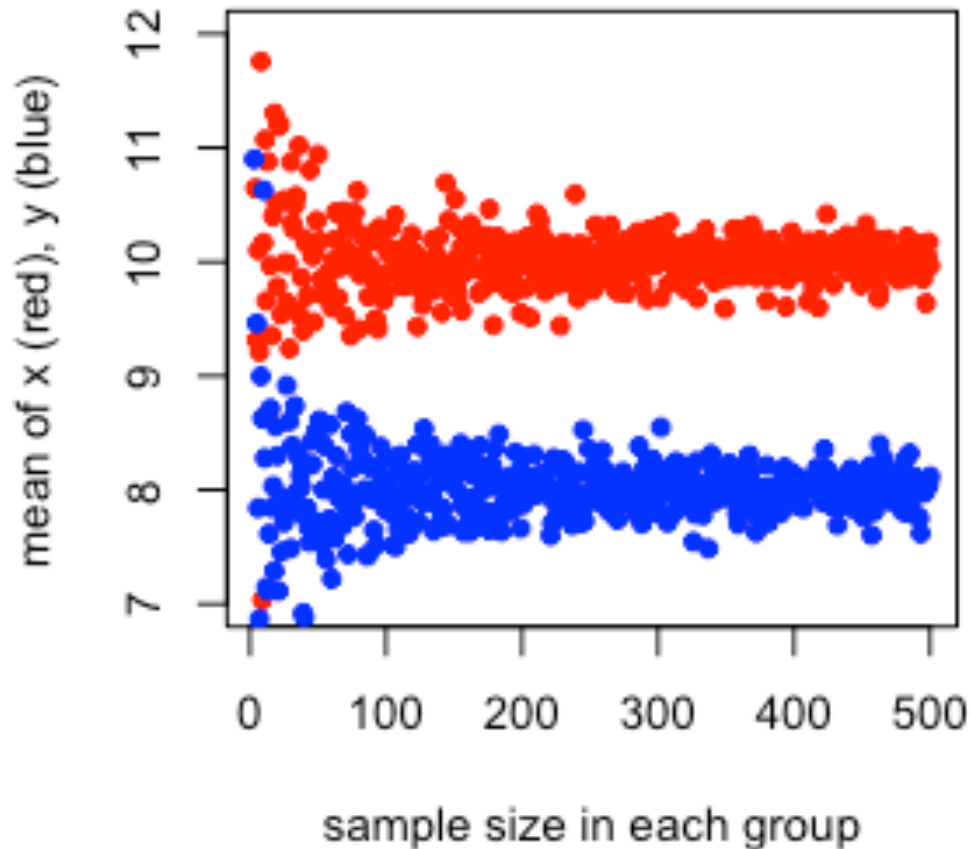
$$diff = \bar{x}_F - \bar{x}_M$$

How about just the difference between means?

- This is a perfectly reasonable measure of effect.
- As long as units are meaningful (and comparable), this can be used. Even if using a standardized measure as well, report this!

$$diff = \bar{x}_F - \bar{x}_M$$

The (unstandardized) difference as a measure of effect



The elephant Vs. mouse tail example

- This “raw” difference can sometimes be challenging to interpret in some contexts..
- Consider the situation of trying to compare sexual dimorphism for tail lengths, in elephants and in mice.
- Why might the raw difference between males and females be challenging to interpret across species?

$$diff = \bar{x}_F - \bar{x}_M$$

Note to self: Do a simulation of elephant and mouse tail SD (first example 20% SD for both, second example 10% for elephants, 30% for mice).

- Show first as unstandardized in cm.

Sexual dimorphism in the Elephant vs mouse tail problem: Multiple solutions

- We will shortly learn about a few ways of scaling our estimates that may be useful.
- Even before this, something as simple as a log transformation of the tail lengths will allow you to focus on proportional differences. Making the comparisons much easier.
- Relatively easy to “back transform” estimates too. No loss of information.

Standardized effect sizes: Using “means”

- Discuss overall mean (whole sample)
- “control group” means
- When will this be useful (when changes in mean phenotype is important, disease studies, genetic studies etc)...

Scaling, standardizing, normalizing

- [https://en.wikipedia.org/wiki/Normalization_\(statistics\)](https://en.wikipedia.org/wiki/Normalization_(statistics))
- Scaling a variable is a generic term, meaning rescaling your variable, but some (usually constant) value. This scaling value could be the mean, sd or something else
- Standardised specifically means scaling by the standard deviation.
 - I can be sloppy and say standarization by the mean. Formally this is incorrect, and I should say scaled by the mean.

Standardized effect sizes: sd based

- Overall pooled SD
- Control group SD (when might you use this)
- When will SD be useful? When the changes you are examining are interesting relative to variation in the population
- Limitations, rubber rulers, estimating SD.

t-test review (two sample t-test):

denominator

- We also need to account for the variation due to sampling (uncertainty).
- How representative would the measure from this class be?
- We capture this using the pooled **standard error** of the mean.
- For this we need the **pooled standard deviation**:

$$s_p = \sqrt{\frac{(n_F - 1)s_F^2 + (n_M - 1)s_M^2}{n_F + n_M - 2}}$$

t-test review (two sample t-test)

- With the denominator being the ***pooled standard error of the mean***.

$$t = \frac{\bar{x}_F - \bar{x}_M}{s_p \sqrt{\frac{1}{n_F} + \frac{1}{n_M}}}$$

$$s_p = \sqrt{\frac{(n_F - 1)s_F^2 + (n_M - 1)s_M^2}{n_F + n_M - 2}}$$

Pooled standard deviation

t-test review

- So the value of t is just the difference in mean heights divided by a measure of (sampling) *uncertainty* in the estimates of mean height for both M and F.

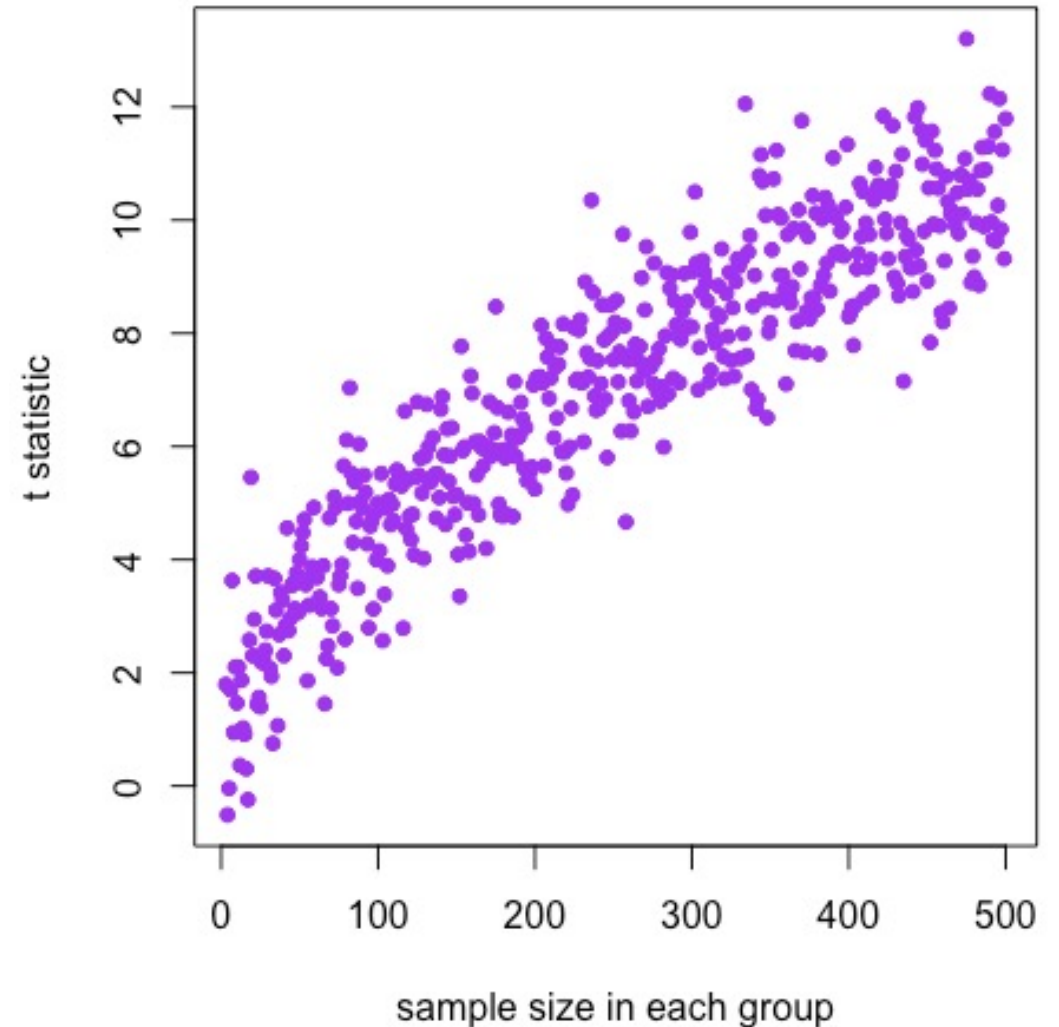
Why is the t statistic NOT a good quantity for an effect size?

$$t = \frac{\bar{x}_F - \bar{x}_M}{s_p \sqrt{\frac{1}{n_F} + \frac{1}{n_M}}}$$

Why is the t statistic NOT a good quantity for an effect size?

$$t = \frac{\bar{x}_F - \bar{x}_M}{s_p \sqrt{\frac{1}{n_F} + \frac{1}{n_M}}}$$

Absolute value of t -statistic will continue to increase with sample size. It does not stabilize. Not appropriate for an effect size.



How might we “standardize” measures

There are two common ways to scale/standardize your measure of effect:

- Scaling by a measure of the “mean” of the trait
 - Scaling by a measure of the “standard deviation” of the trait
- Both of these approaches can be useful in some situations, and which you choose will depend on your goals (and your field).

How about standardized effect sizes: Let's start with *Cohen's d*

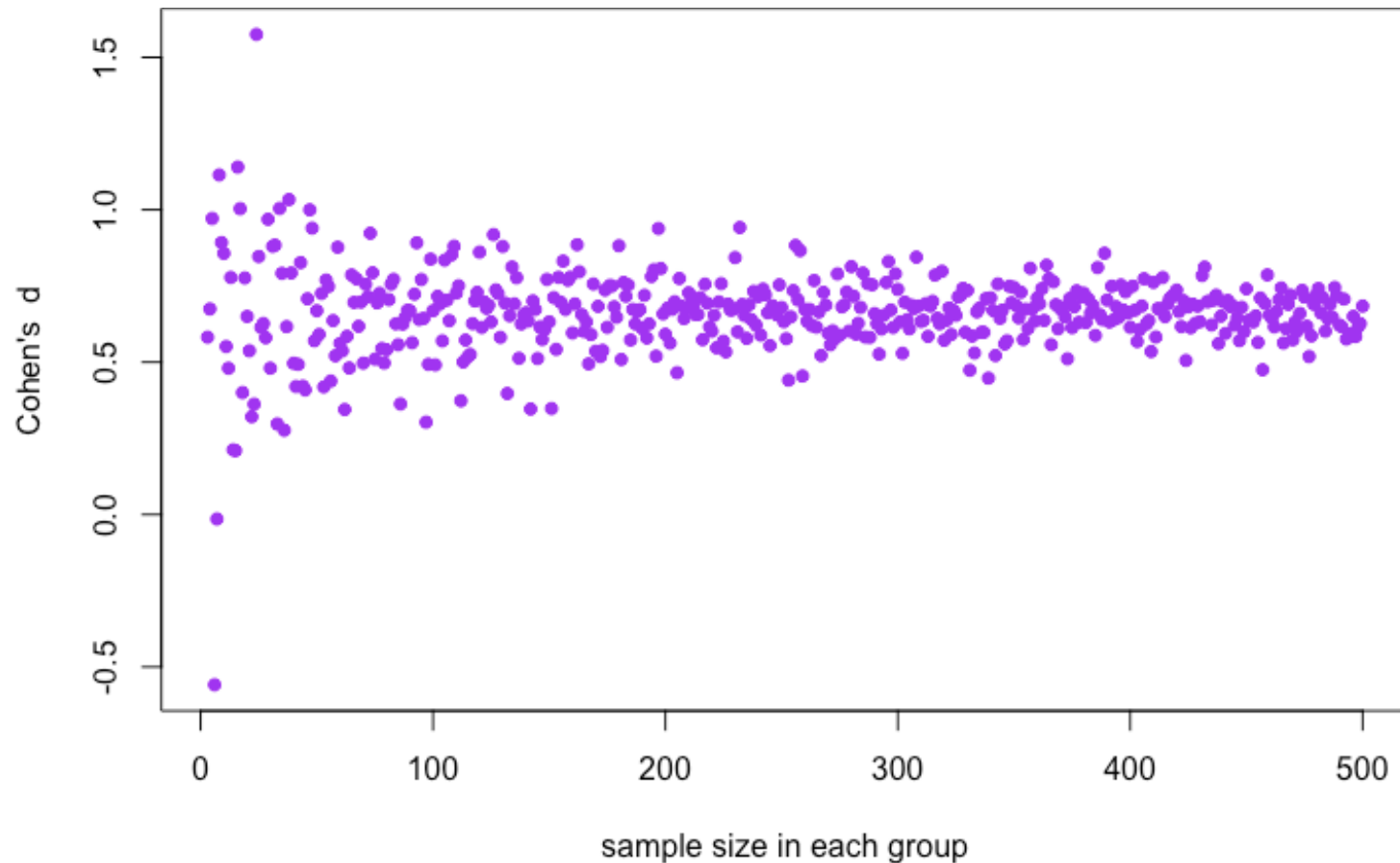
$$\text{Cohen's } d = \frac{\bar{x}_F - \bar{x}_M}{S_{pooled}}$$

$$s_p = \sqrt{\frac{(n_F - 1)s_F^2 + (n_M - 1)s_M^2}{n_F + n_M - 2}}$$

The difference just scaled by the pooled standard deviation. As a reminder this serves as a ***measure of biological variation***, not uncertainty in our estimates like the standard error.

How about effect sizes: Let's start with *Cohen's d*

$$Cohen's\ d = \frac{\bar{x}_F - \bar{x}_M}{S_{pooled}}$$

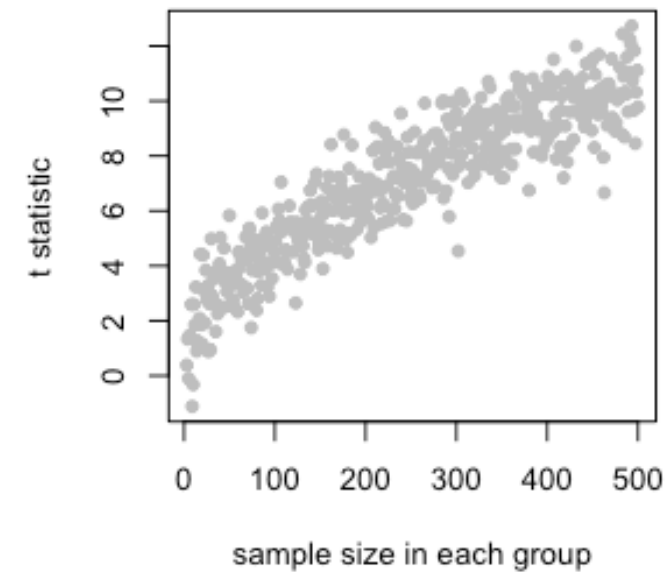
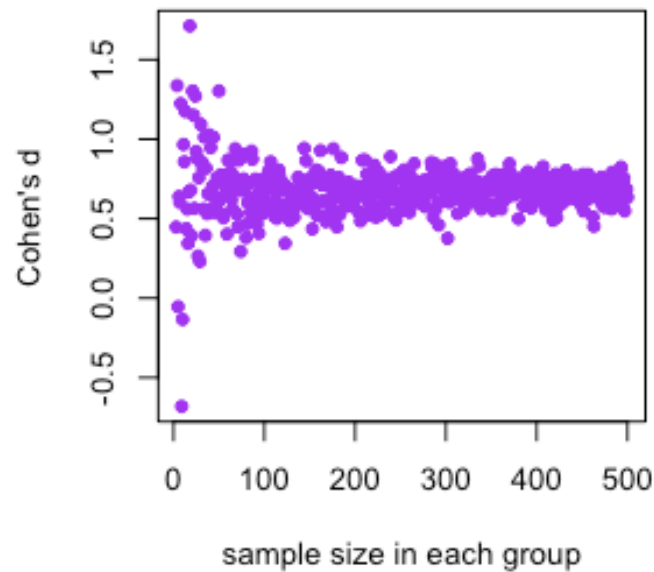
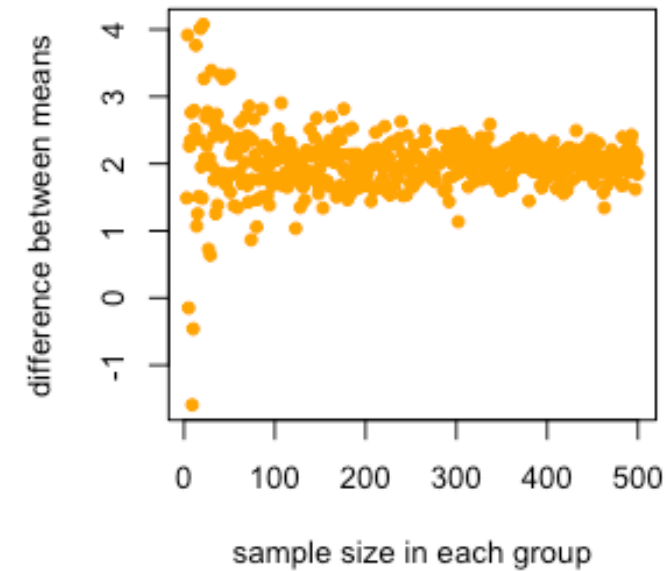
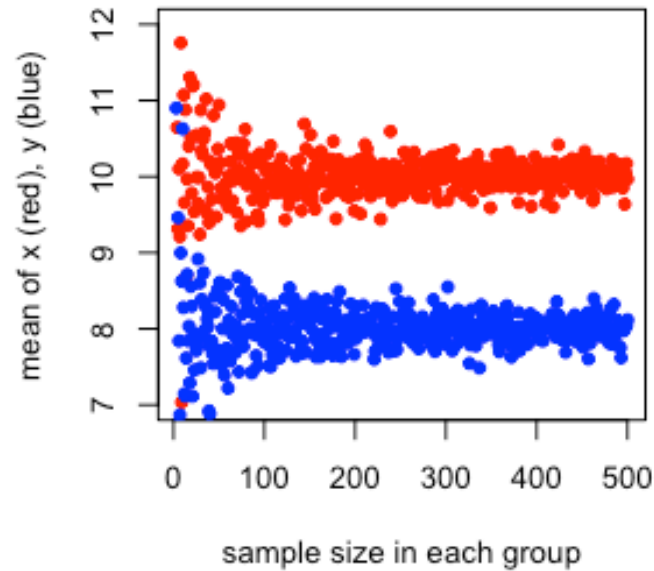


Cohen's d shows the same overall pattern, just gets more precise with n .

Think about it this way:

If you are comparing a measure of the difference of the mean of x and y , what is most consistent with your goals of estimating a meaningful measure of effects?

So using the *difference*, or *Cohen's d* are sensible for effect sizes. Not *t*!

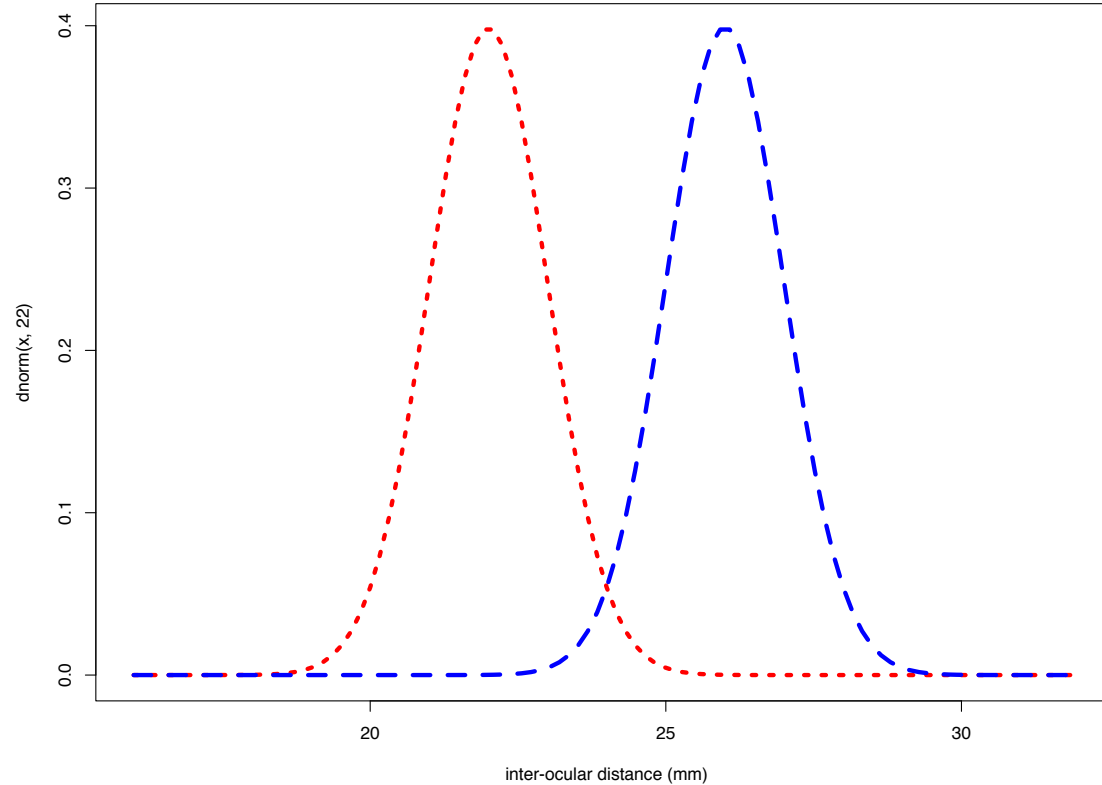


Interpreting Cohen's d: thinking in terms of standard deviations...

$$\text{Cohen's } d = \frac{\bar{x}_F - \bar{x}_M}{S_{pooled}}$$

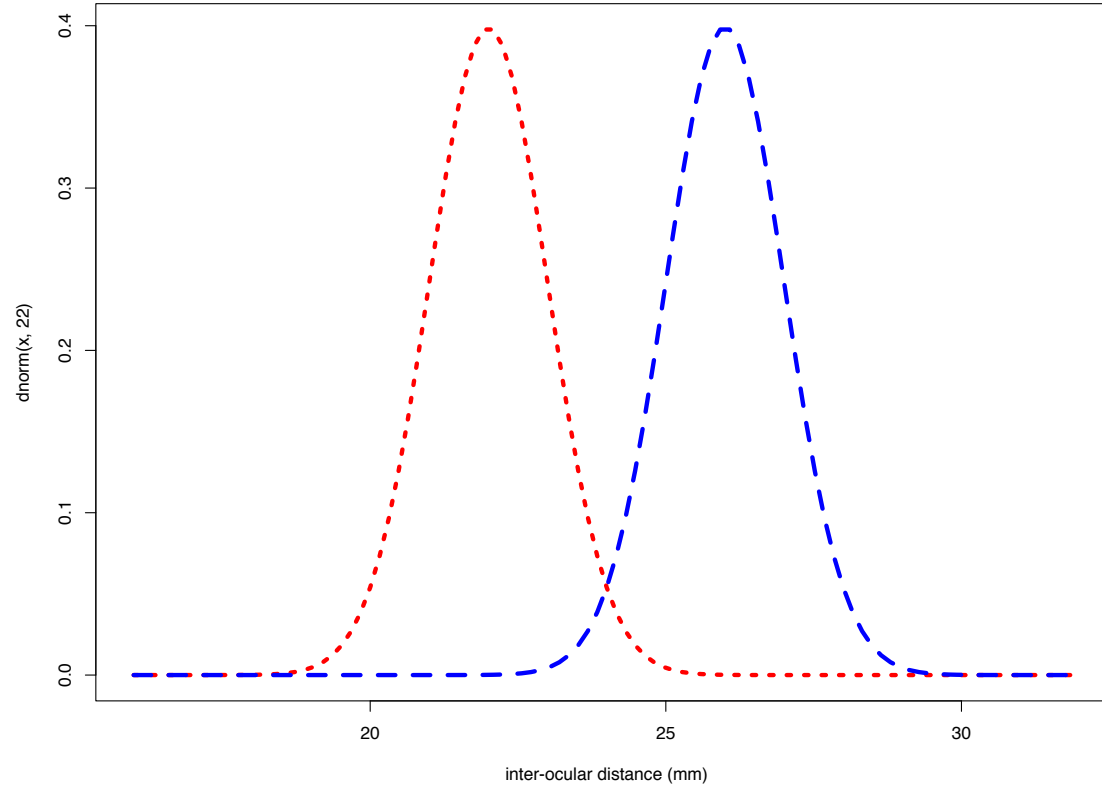
(maybe add visualization for distributions of each group relative to difference...)

Why do I keep going on about “pooled”



What would happen to the estimate of the variance/sd if I just combined all of the data together, ignoring the mean?

Why do I keep going on about “pooled”



Each group has a $sd = 1$,

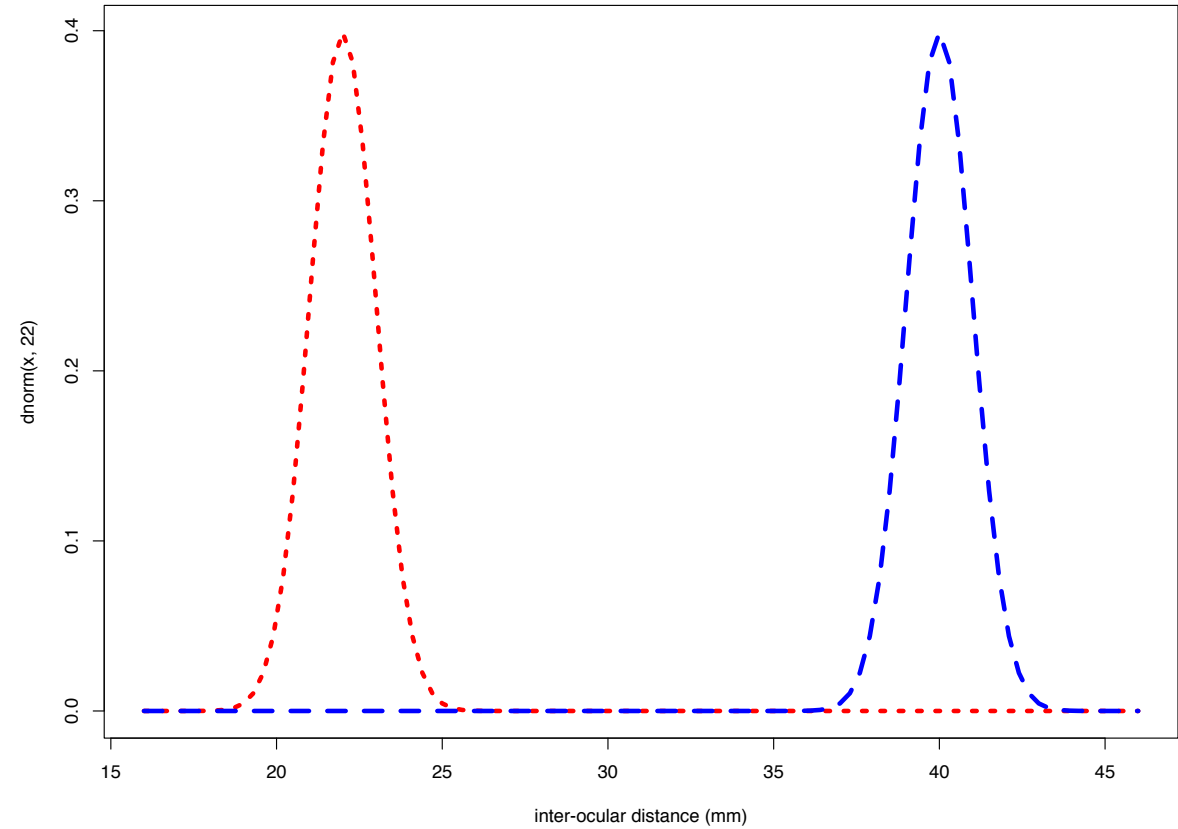
If I just lump them together
(ignoring red and blue) they have
a \sim sd of 2.23

!Show an example simulation of what happens when Spooled is used, or a naïve SD!

Even further apart

Each group has a sd = 1,

If I just lump them together... they now have a sd of ~9




What would happen to our estimate of *Cohen's d* if we did not do proper pooling of our *sd*?

$$\text{Cohen's } d = \frac{\bar{x}_F - \bar{x}_M}{S_{total}}$$

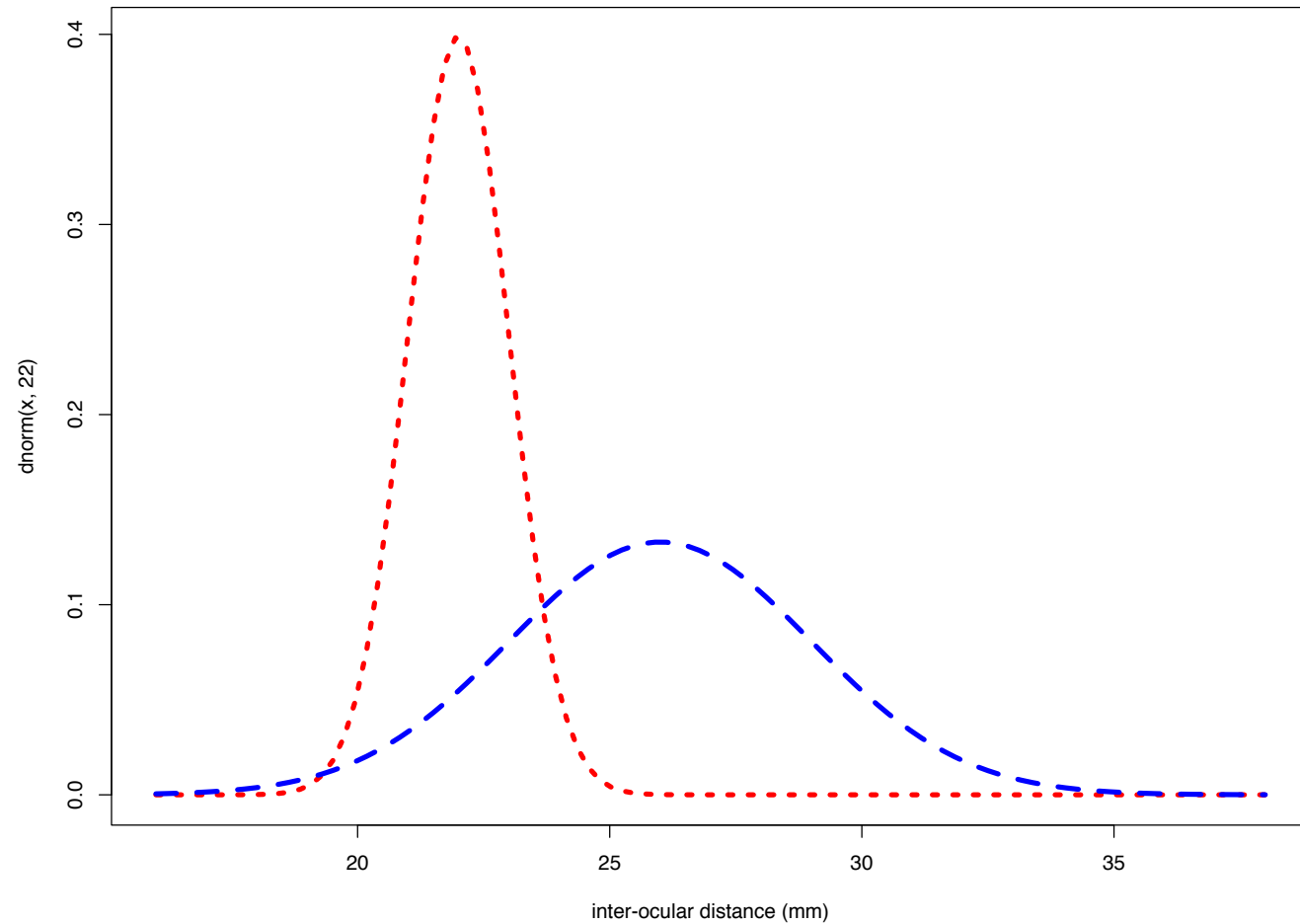
...It would cause us to underestimate the magnitude of the difference

What would happen to our estimate of *Cohen's d* if we did not do proper pooling of our *sd*?

$$\text{Cohen's } d = \frac{\bar{x}_F - \bar{x}_M}{S_{total}}$$


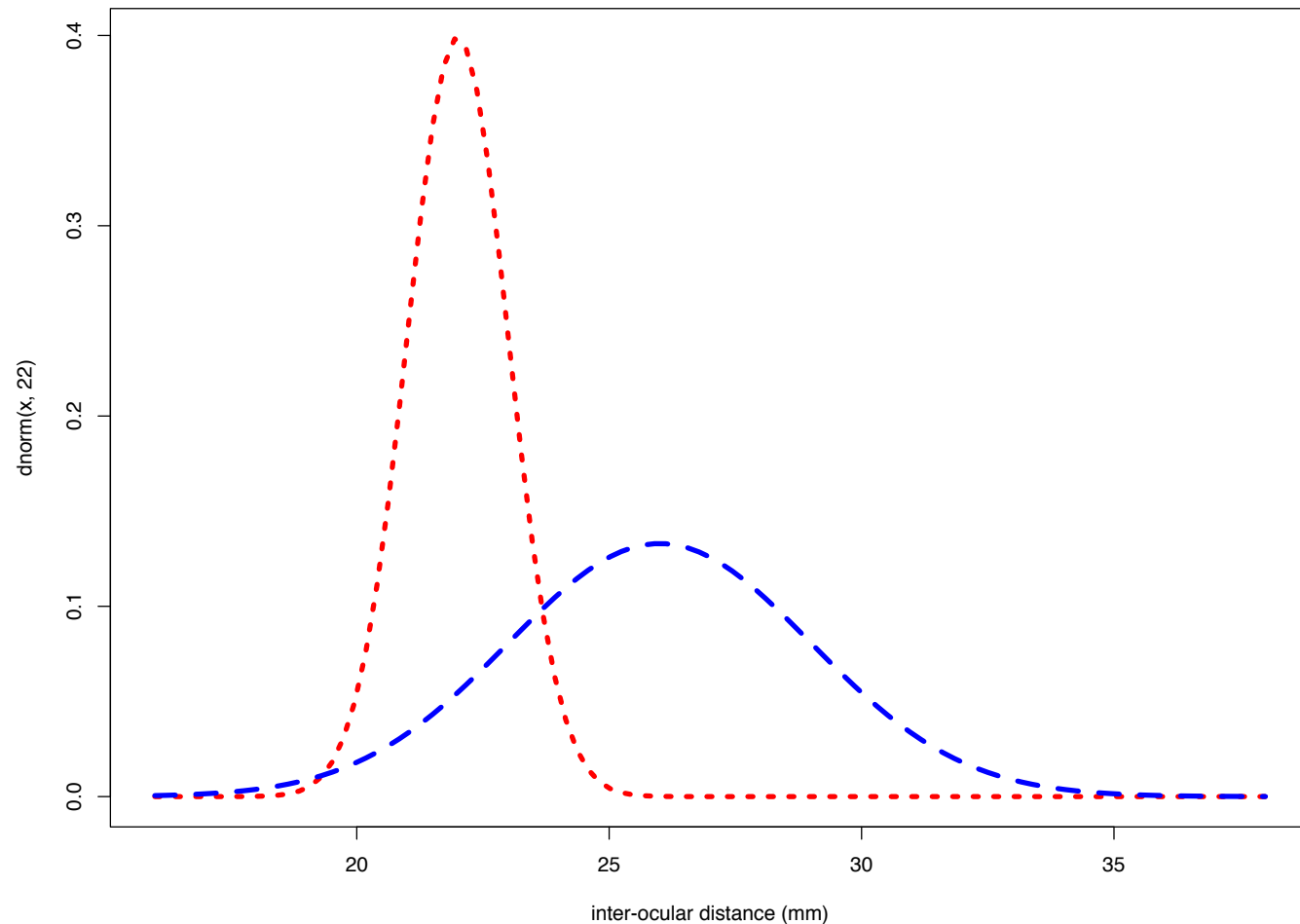
...It would cause us to underestimate the magnitude of the difference

Also this.... (heterogeneity in the variability across groups, also known as heteroscedasticity)



Variances may not be equal in each group....

Heteroscedasticity causes its own issues for this...
We can discuss how to deal with this when we learn
how to use the non-parametric bootstrap



A few measures of effect size.

Table 1. Equations for calculating d statistics

Case	Equation	Description	References
Comparing two independent or dependent groups (i.e. both paired and unpaired t -test cases)	$d = \frac{m_2 - m_1}{s_{\text{pooled}}}$	(1) m_1 and m_2 are means of two groups or treatments, s_{pooled} is pooled standard deviation,	Cohen (1988); Hedges (1981)
	$s_{\text{pooled}} = \sqrt{\frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2}{n_1 + n_2 - 2}}$	(2) n is sample size (in the case of dependent design, the number of data points), s^2 is variance.	
Comparing two independent groups (i.e. unpaired t -test case)	$d = t_{\text{unpaired}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$	(3) Alternatively, t values can be used to calculate d values; t_{unpaired} is the t value from the unpaired t -test (compare with Equation 10 in the text)	Rosenthal (1994)
Comparing two dependent groups (i.e. paired, or repeated-measure t -test case)	$d = t_{\text{paired}} \sqrt{\frac{2(1 - r_{12})}{n}}$	(4) t_{paired} is the t score from the paired t -test, r_{12} is correlation coefficient between two groups, and note that $n = n_1 = n_2$ not $n = n_1 + n_2$	Dunlap <i>et al.</i> (1996)

Free software by David B. Wilson to calculate these effect statistics is downloadable (see Table 4). Strictly speaking, Equations 1 to 4 are for Hedges's g but in the literature these formulae are often referred to as d or Cohen's d while Equation 10 is Cohen's d (see Kline, 2004, p.102 for more details; see also Rosenthal, 1994; Cortina & Nouri, 2000).

Good news... easy to convert from t to d or r

$$d = \frac{t(n_1 + n_2)}{\sqrt{n_1 n_2} \sqrt{df}}$$

$$r = \frac{t}{\sqrt{t^2 + df}}$$

n_1 and n_2 are sample sizes for groups 1 and 2.
 df are the degrees of freedom (from t).

Unbiased estimate of d (sometimes called *Hedges' d*)

- As mentioned in your readings, for small sample sizes (less than 25 or 30 total samples) d can be upwardly biased (too big).
- Here is a common correction. For larger sample sizes these corrections tend not to matter.

$$d_{unbiased} = \left(1 - \frac{3}{4 \times df - 1}\right) \times d$$

Unbiased estimate of d , (sometimes called Hedges' d or g)

- As mentioned in your readings, for small sample sizes (less than 25 or 30 total samples) d can be upwardly biased (too big).
- Here is a common correction. For large sample sizes these tend not to matter.

$$d_{unbiased} = \left(1 - \frac{3}{4 \times df - 1}\right) \times d$$

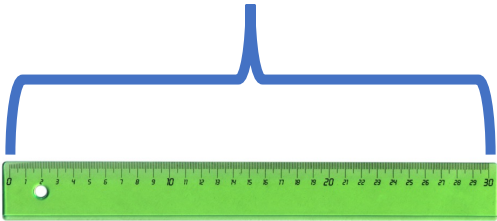
For single-group or paired designs $df = (n-1)$

For two independent groups using s_p as the denominator in d , then $df = (n_1 + n_2 - 2)$

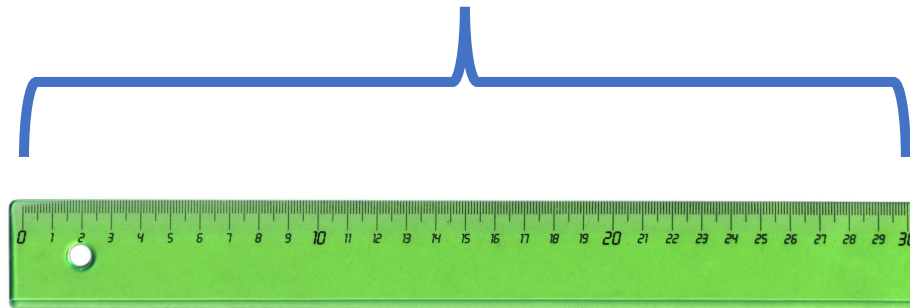
Potential issues about scaling by standard deviation?

”Rubber ruler” effect

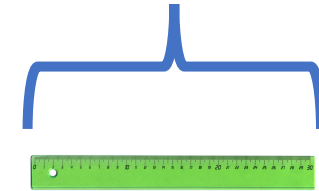
1 S.D. in study 1



1 S.D. in study 2



1 S.D. in study 3



- We may not always be scaling by the same value if we always determine standard deviation within study.
- As the square root of the second moment, estimating it well takes more data compared to the mean (first moment).

Can you think of other ways to scale the measures of effect sizes?

- Scale by the $sd_{control}$ group (*Glass's delta*)
- Scale by the mean (either $mean_{pooled}$ or $mean_{control}$).
 - In quantitative genetics we use both heritability ($h^2 = \frac{V_A}{V_P}$) and coefficient of genetic variation ($CV_A = \frac{V_A}{\mu^2}$) as ways to compare genetic variation among populations.
- You just need to spend some time to figure out what they mean with your data and whether they are sensib.e

other advantages with using effect sizes

- If effect sizes and CI are published you can always calculate p values.
- Effect sizes + CI are useful for meta-analyses.

Pearson correlation coefficient (r)

- Why is this a meaningful measure of effect size?

Concerns about standardized effect sizes

- ***Rubber ruler effect***: the denominator (along with the numerator) is also estimated from the data, so the standardization changes for each sample as well. This can be problematic (for small sample sizes in particular).
- **NOTE FOR ID: simulate with only *sd* or only *diff* varying to demonstrate.**
- In particular variances (and thus *sd*) are second moments, so are even more data hungry to estimate well.
- Can contrast difference in original units to standardized measure.
- If you have lots of values in the literature, can use a constant *sd* that is pre-determined or an empirically derived prior.

What might we do if we have many levels to a given categorical predictor?

- We can use the estimated variance component for that predictor.
- It can be expressed as the sqrt of the variance component to place it in units of the response.
- It can be thought of in relation to the observed variance for the response.
- Scaling it $V_{\text{Treatment}}/V_{\text{observed}}$ may be useful. i.e. heritability.
- You could also scale by the mean (coefficient of variation)

Confidence Intervals

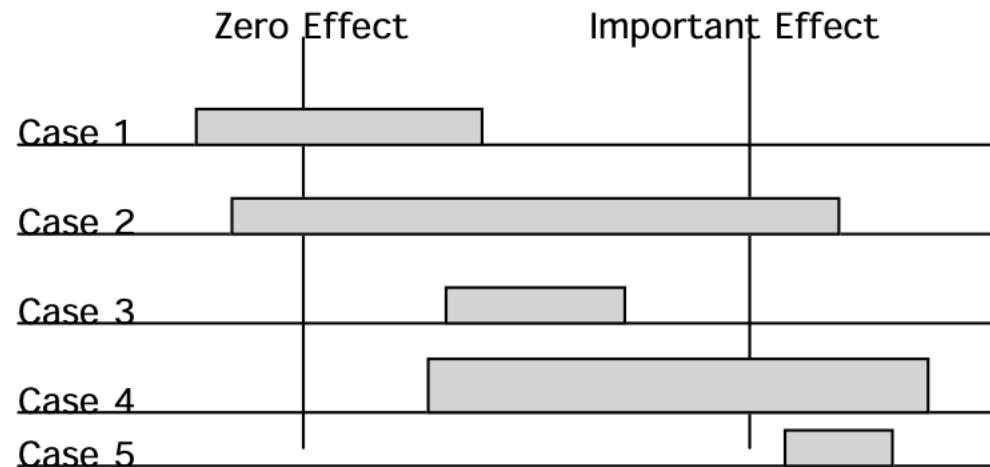
- For confidence intervals for any of these I recommend using monte carlo methods, resampling or Bayesian (MCMC) approaches to derive the CI.
- You can even using these methods to generate CI for R^2 (we will do some examples with Monte Carlo and non-parametric bootstraps).

How big does an effect need to be for it to be biologically meaningful?

- Sorry, I can't answer that for you.
- This will depend a lot on the field you are in and the biology of the system.
- Some authors suggest particular thresholds for *Cohen's d*. I don't buy it as a general tool (No different than an arbitrary alpha).

So for example

- Based on what you know about the biology of your system you could *a priori* (really really critical) decide on something like the following...

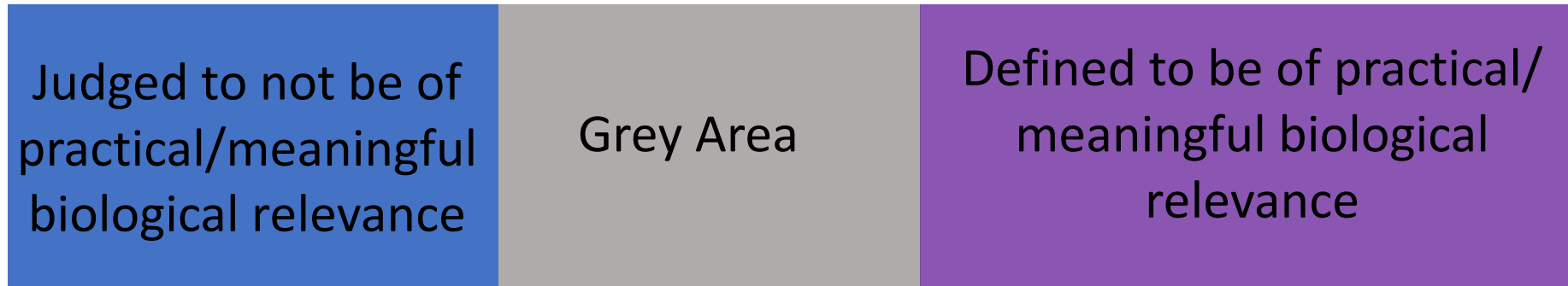


Fox 2001,
Environmetrics 12: 437-449

Figure 1. Confidence intervals for five different environmental scenarios.

Measure of effect -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0

a priori



Measure of effect -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0

a priori

Judged to not be of practical/meaningful biological relevance

Grey Area

Weak effect

Moderate effect

Strong effect

Pre-defined to be of practical/meaningful biological relevance

- It is good (great!) if you are able to distinguish magnitudes of weak, moderate and strong effects if possible.
- But don't make absolute thresholds with this approach either (use the information of the entire interval to make inferences). Otherwise, you are substituting one problem (thresholds based on p-values and point null values) for another (threshold based on magnitudes).
- See <https://easystats.github.io/effectsize/articles/interpret.html>

When Benchmarking, small, medium, large, etc, remember:

”Admittedly, if people interpreted effect sizes with the same rigidity with which $\alpha = 0.05$ is used in statistical (*significance*) testing, we would merely be being stupid in another metric.”

You perform your experiments, and these are your results.. What do they mean?

Measure of effect -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7 .8

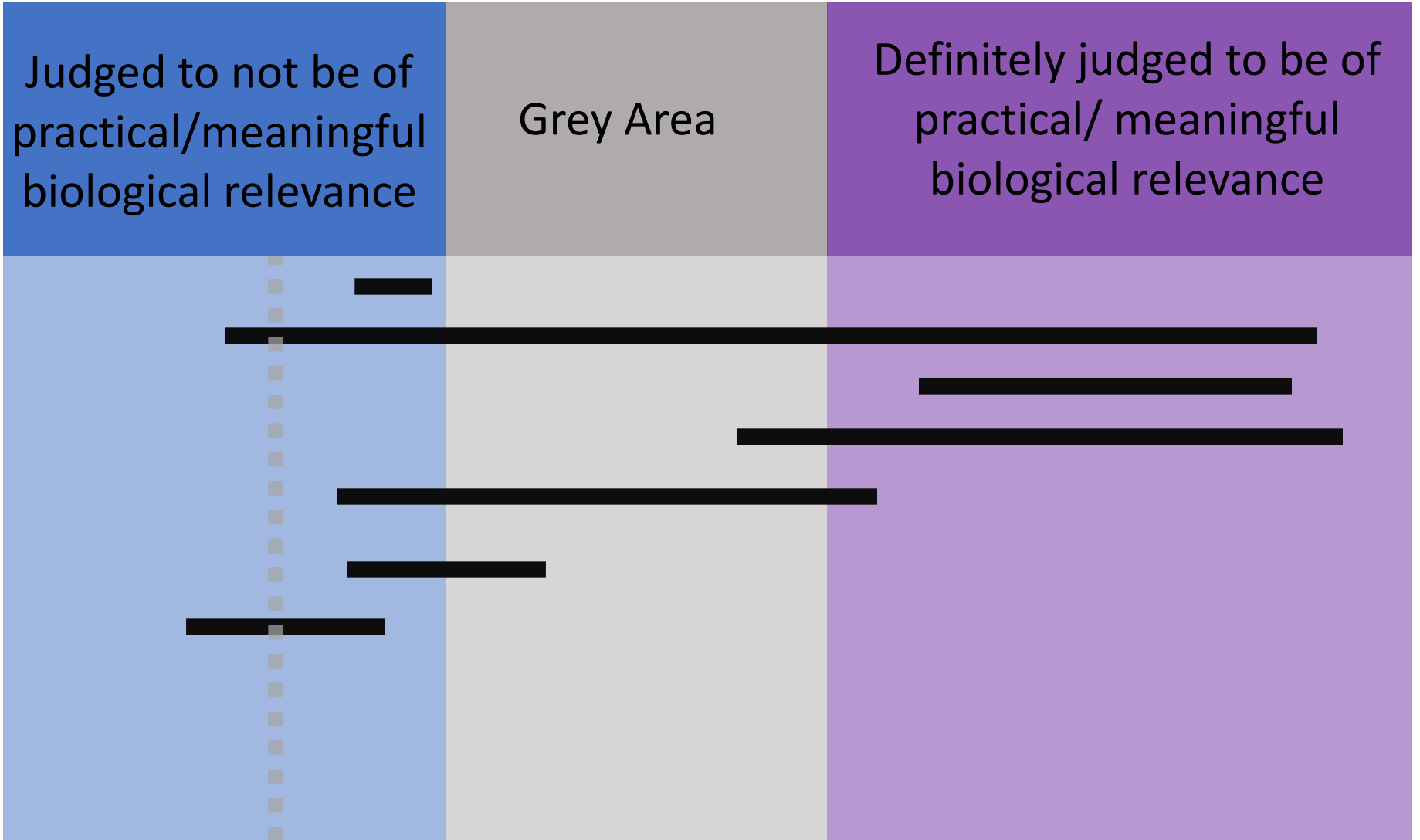
a priori

Judged to not be of practical/meaningful biological relevance

Grey Area

Definitely judged to be of practical/ meaningful biological relevance

Observed
"results"
from
experiments



1
2
3
4
5
6
7

R^2 : The co-efficient of determination

- *Not really a measure of effect, but often used as such.*
- R^2 is probably the most commonly used quantity for model fit.
- Often described as the proportion of variation *explained* by the model.
- I prefer: **proportion of variation *accounted for* by the model.**
- $1 - R^2$: proportion of variation not accounted for by model

R^2 : The co-efficient of determination

- $SS.total = SS.model + SS.residual$
- (un-adjusted) $R^2 = 1 - (SS.residual/SS.Total)$
= $SS.model/SS.Total$
- $0 \leq R^2 \leq 1$
- However, when you add more parameters to a model, at worse they do not increase $SS.model$ (they will never decrease it).
- Effectively unadjusted R^2 will always increase with more parameters added to the model.
- It does not penalize more complex models (violating our parsimony principal).

Adjusted R^2

- Adjust for parsimony principle
- Adjusted $R^2 = 1 - (n-1)/(n-p)(1-R^2)$
= 1 - residual MS/total MS
- Adj. R^2 can decrease with increasing numbers of parameters (p).
- Information theoretic approaches are still far better ways of comparing different models.

Is R^2 useful?

- It is useful in making a statement about **overall** model fit (% variation accounted for).
- It is ***not useful*** in comparison between models.

Model vs predictor specific R^2

- While in a glm with multiple predictors, the coefficients are adjusted for the presence of one another, this is not the case for R^2 .
- Most statistical software provides the R^2 for the full model.
- So how do we assess variance accounted for at a predictor level?

Coefficient of Partial Determination

Partial R^2

- We can instead adjust the R^2 in a manner analogous to adjusting coefficients for other predictor variables.
- These are called partial R^2 (named to provide similar meaning to partial regression coefficients.).
- These allow you to adjust the R^2 for a given predictor, given all of the other predictors in the model.
- You can do this in R using the `partial.R2` function in the `asbio` library.

```
partial.R2(model.without.predictor, model.with.predictor)
```

Salient Points

- There are several classes of effect sizes (unstandardized, scaled by a measure of the mean, scaled by pooled sd , variance accounted for, odds ratios).
- Deciding which to use depends on the question at hand, and with what you compare your results to.
 - Always useful to include the unstandardized measure.
- While deciding what to use can take time and thought, it will likely save you time when you are interpreting your model estimates.

Salient points of the material

- Do not feel obliged to use just these. If there are other sensible measures that aid in the interpretation of your results, use them.
- Also do not feel like you can only use one. Examining different measures of effect sizes may help you understand what the model is telling you!

Table 3. Asymptotic estimates of standard errors (se) and other formulae required to calculate confidence intervals

Statistic	Equation	Note	References
d (independent, unpaired)	$se_d = \sqrt{\frac{(n_1 + n_2 - 1)}{(n_1 + n_2 - 3)} \left[\left(\frac{4}{n_1 + n_2} \right) \left(1 + \frac{d^2}{8} \right) \right]}$ (16)	Equation 16 provides se for Cohen's d while Equation 17 provides se for Hedges' d (unbiased d in Equation 14)	Hunter & Schmidt (2004); Hedges (1981)
	$se_d = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}}$ (17)		
d (dependent, paired, repeated measure)	$se_d = \sqrt{\frac{2(1 - r_{12})}{n} + \frac{d^2}{2(n - 1)}}$ (18)	$n = n_1 = n_2$, and r_{12} is correlation coefficient between two groups	Becker (1988)
r (correlation coefficient)	$se_{\mathcal{Z}r} = \frac{1}{\sqrt{n - 3}}$ (19)	$\mathcal{Z}r$ is the Fisher transformation of r and the distribution of r is not normal but that of $\mathcal{Z}r$ is normal	Hedges & Olkin (1985)
	$\mathcal{Z}r = 0.5 \ln \left[\frac{(1 + r)}{(1 - r)} \right]$ (20)		
	$r = \frac{e^{2\mathcal{Z}r} - 1}{e^{2\mathcal{Z}r} + 1}$ (21)		

Refer to Table 1 for some of the symbols used.