Biological VS statistical significance

date

Readings for effect sizes

- See the PDFs in the ANGEL folder

Readings for this Thursday (review of the linear model)

Review readings (if nesc):

Dalgaard: Chapters 6,7,11,12

Main readings:

covering all of linear models in about 2-3 lectures

GelmanHill: Chapters 3, 4,Appendix A

R_intro_guide chapter 11 (a lot of useful advice for R syntax for using lm()).

Also helpful for people who want a more mathematical treatment:

Faraway, J. (Linear Models in R) Chapter 2

(His free online book (available in the books folder on ANGEL) also covers similar material.

Readings for Probability (For next Thursday).

- We start our field guide to probability.
- For review, Dalgaard Chapters 3-4

- Primary readings Bolker Chapter 4. Gelman and Hill Chapter 2 (pages 13-26).

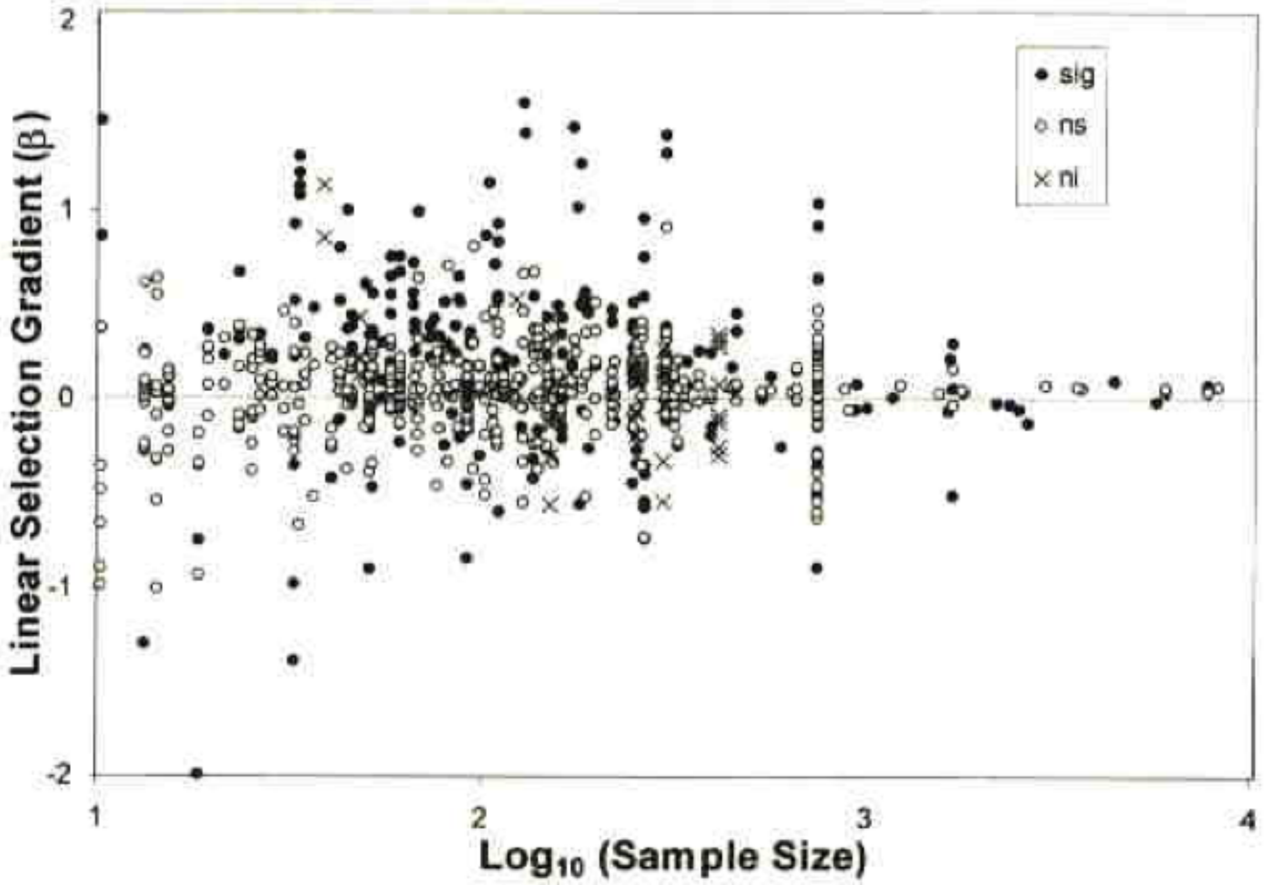
- For more advanced readings in probability see syllabus

IGNORE FOR NOW

Goals for the day

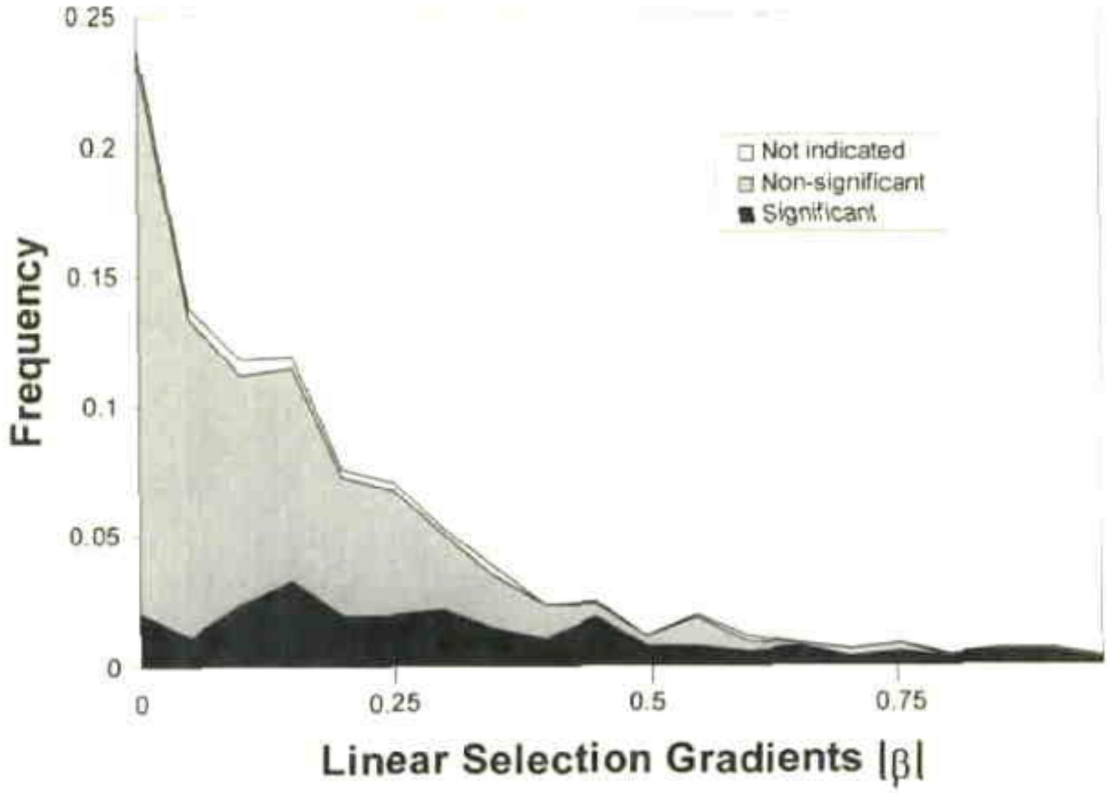
Discuss the idea of effect sizes, and why (with CIs) they are central to any statistical and biological inference

Motivating example: The strength of natural selection in the wild



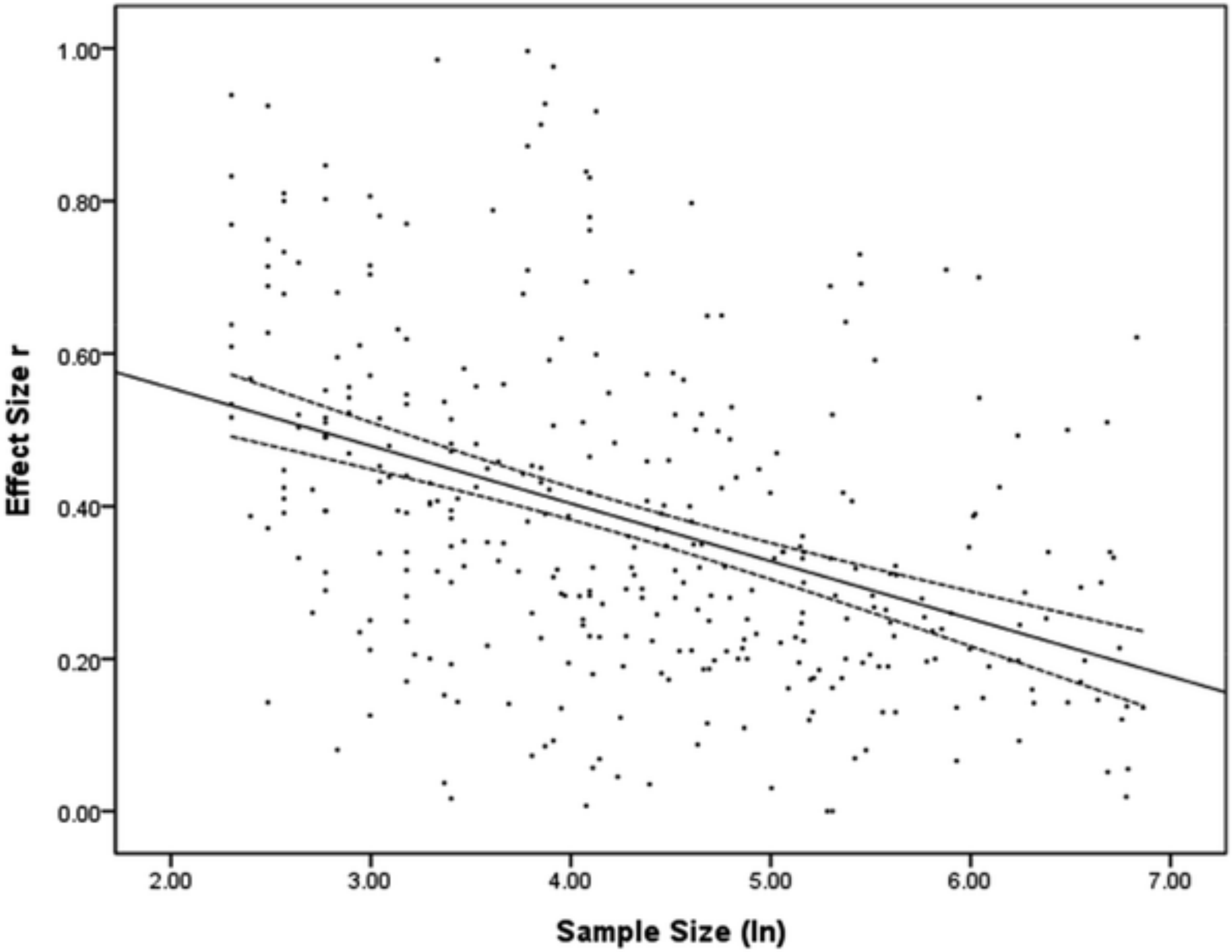
Kingsolver et al 2001

Is directional selection strong in nature?



Kingsolver et al 2001

Figure 5. Corrected effect size r plotted against logarithmically transformed sample size.



Kühberger A, Fritz A, Scherndl T (2014) Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. PLoS ONE 9(9): e105825. doi:10.1371/journal.pone.0105825
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0105825

The Big Picture

- The coefficients (estimated parameters) from our models are not simply estimates to be examined along with p-values, but are probably the most important aspect of the model with respect to your ability to assess the importance of particular variables.

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 one to use may depend a fair bit on the question at hand, and what you plan to compare your results to.
- This can take a considerable amount of thought.

What problems can we have with a null hypothesis i.e. $|\beta_0| \leq x$?

two things: 1) we assume the null=0 when really this value can be biologically unrealistic (height of zero)
2) with large enough sample sizes, we can reject any null hypothesis. As n increases, power increases, and we can begin to have the sensitivity to detect really minor (and generally non biologically meaningful) trends.

The problem with a point null of $\beta_0=0$

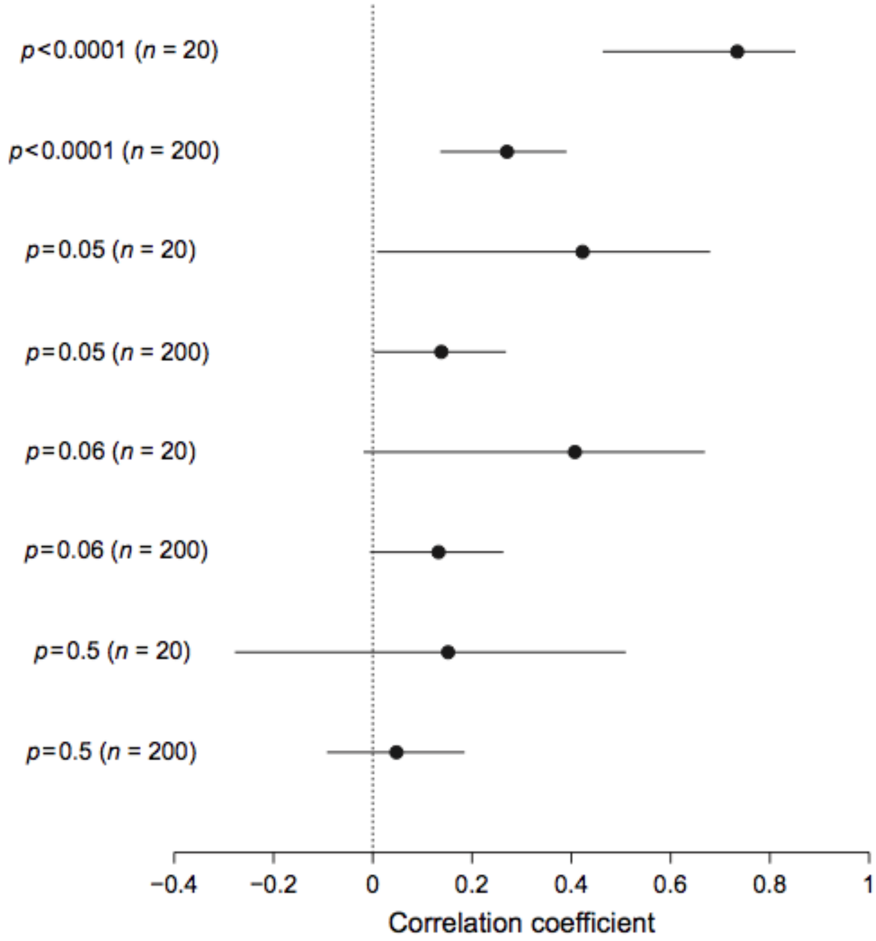
in biological systems, ß tends to move because of variation, sampling/measurement error

perhaps null should be made relative to known measurement error (minimum limit on measurement of certain tool, for instance)

Publication bias

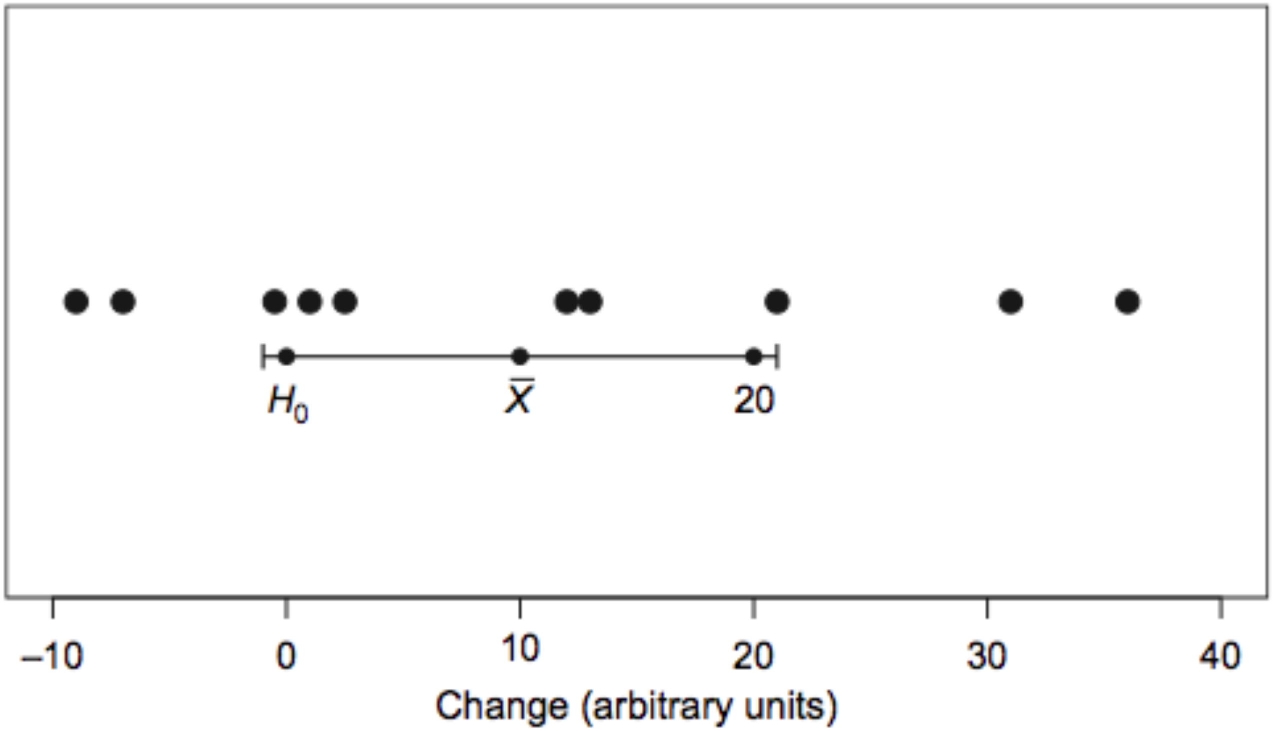
- There is a huge literature on publication bias towards “significant results”.
- That is many studies have test statistics just a bit less than the magic alpha.

Practical VS statistical significance



obviously many of these comparisons where p = 0.05 vs. 0.06 do not have huge differences! would you be excited about one and not the other? no fucking way. Here we also see that increasing n has a huge effect on CIs. Let’s use CIs to detect exclusion of mean differences of zero

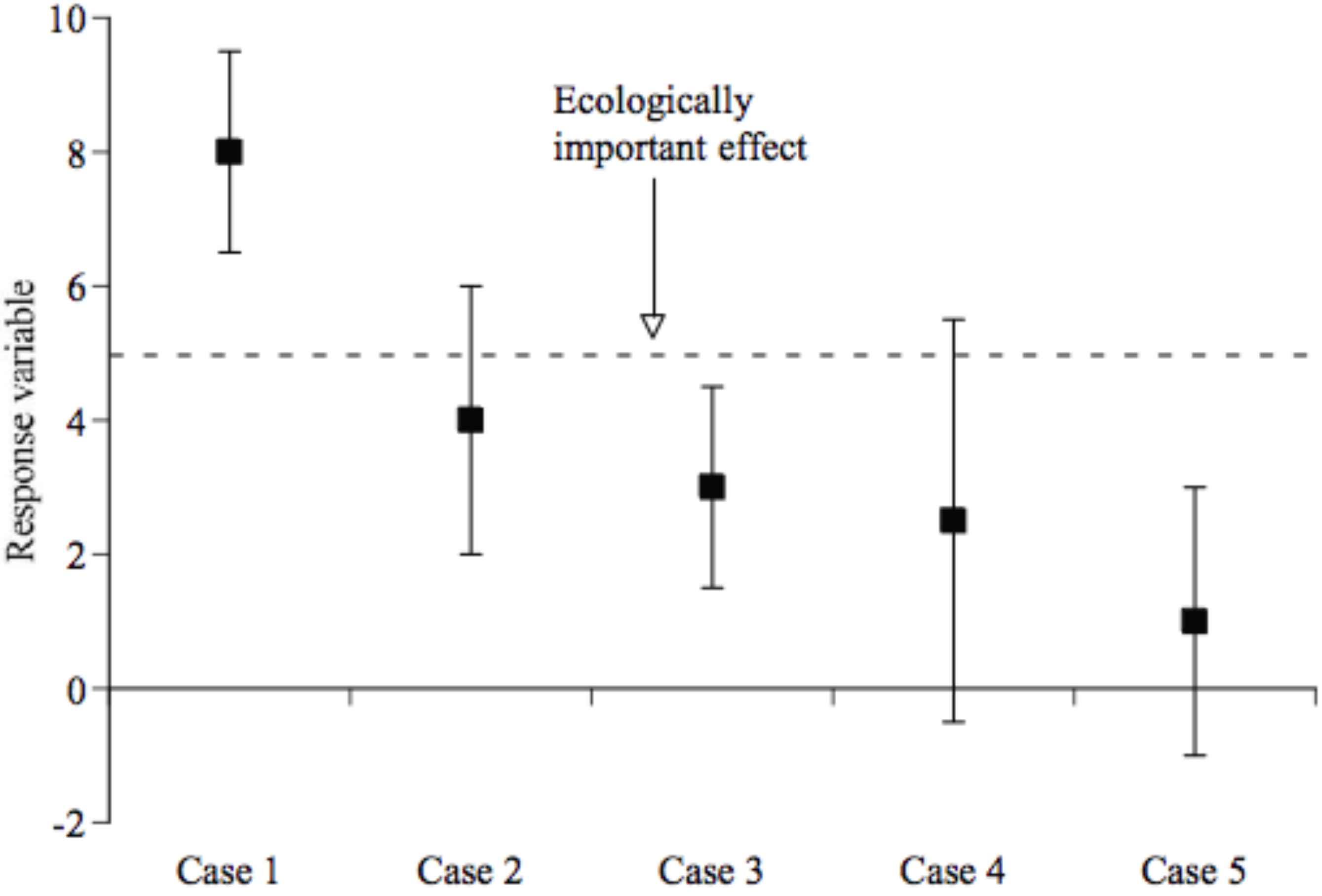
The Counter-null



Nakagawa & Cuthill 2007

We always focus on whether the CI includes zero in determining significance, but what about the other side of the CI? these values are just as important, and they're often disregarded!! The counter-null

It may be significant, but is it important?



A few measures of effect size.

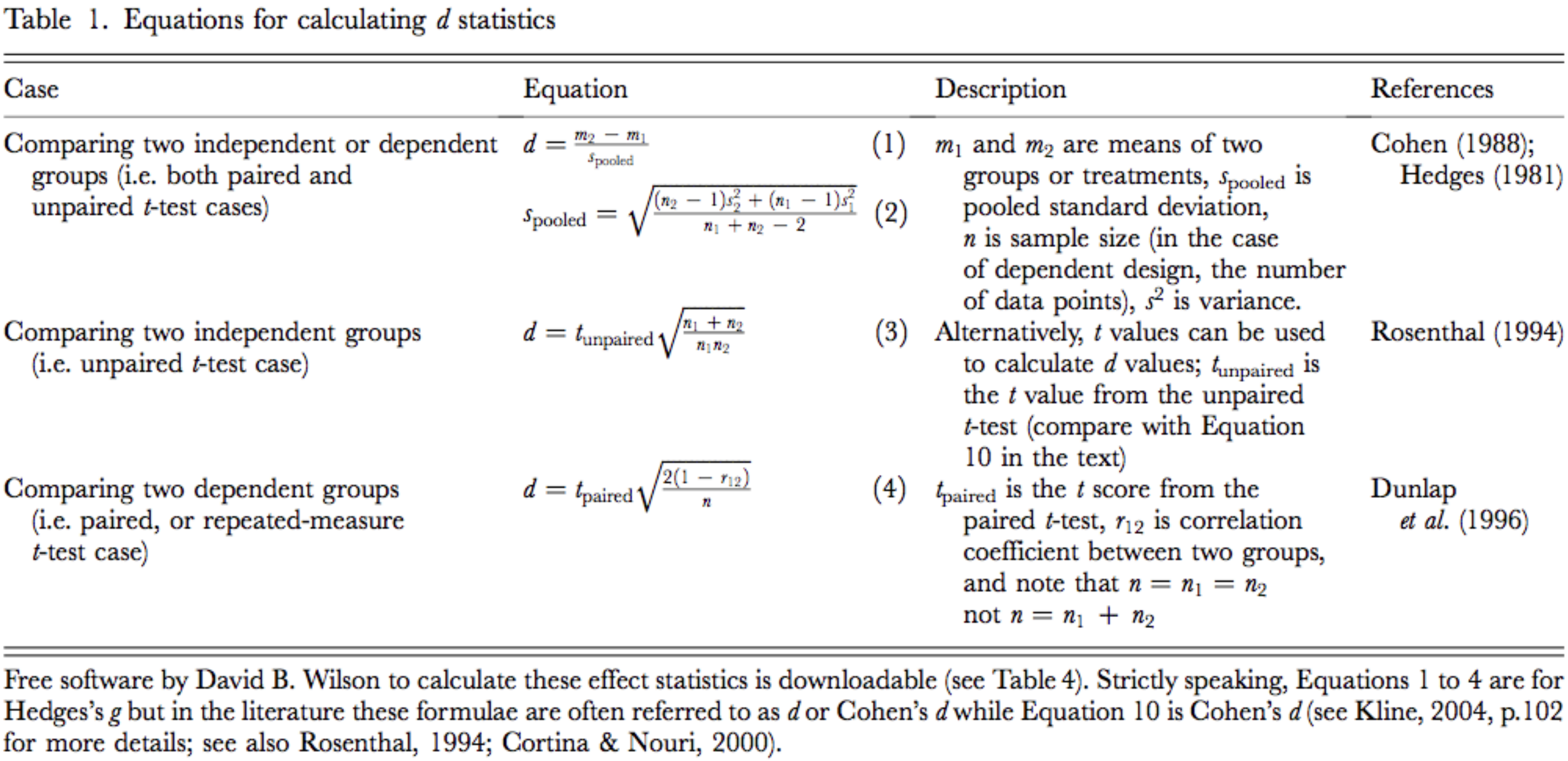


Table 1. Equations for calculating *d* statistics

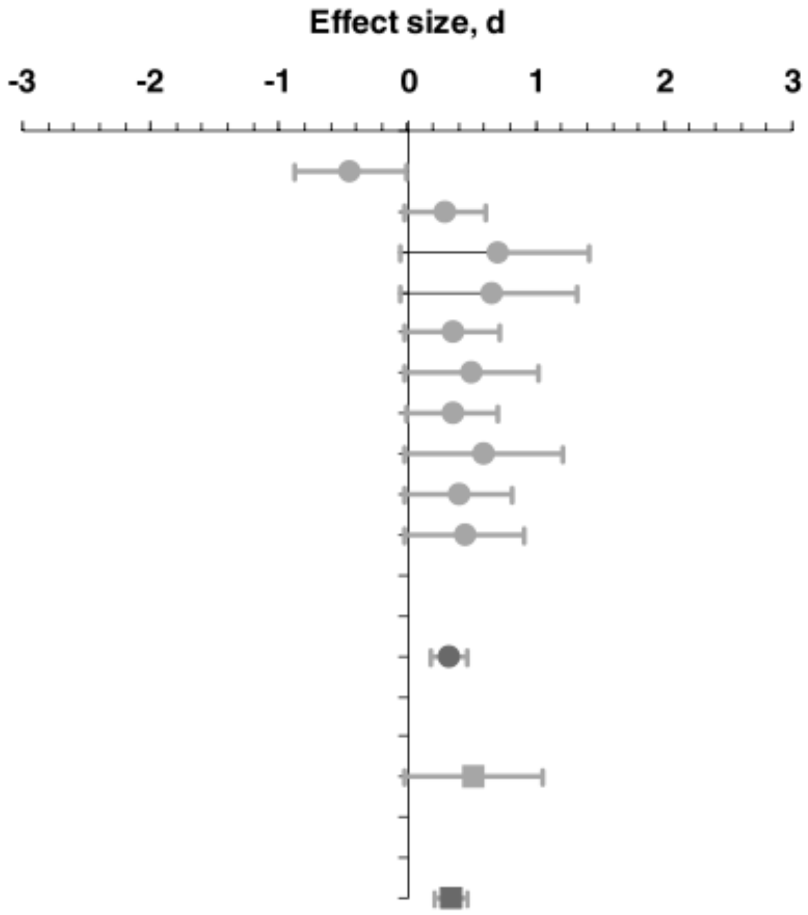
Case	Equation	Description	References
Comparing two independent or dependent groups (i.e. both paired and unpaired *t*-test cases)	$d = \frac{m_2 - m_1}{s_{\text{pooled}}}$ (1) $s_{\text{pooled}} = \sqrt{\frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2}{n_1 + n_2 - 2}}$ (2)	m_1 and m_2 are means of two groups or treatments, s_{pooled} is pooled standard deviation, n is sample size (in the case of dependent design, the number of data points), s^2 is variance.	Cohen (1988); Hedges (1981)
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 *et al.* (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).

don't worry about the software; it's all trivial to code in R

Nakagawa & Cuthill 2007

Effect size in context



Thompson 2007

What other advantages

- If effect sizes and Cis are published you can always calculate p values.
- The converse is not true from a p value and a test statistic with dfs.
- Effect sizes + CIs. are much more useful for meta-analyses.

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

- Scale by the $\text{sd}_{\text{control}}$ group
- Scale by the mean (either $\text{mean}_{\text{pooled}}$ or $\text{mean}_{\text{control}}$).

- However for each of these you do need to spend some time to figure out what they mean!

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 $V_{Treatment}/V_{observed}$ may be useful. i.e. heritability.
- You could also scale by the mean (coefficient of variation for the)

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).

Interesting new idea (for both hypothesis testing and assessing importance of effect).

- “Protected Inference”
- If your effect size is smaller than measurement error, perhaps you need to consider (even if formally significant) whether it is of use.

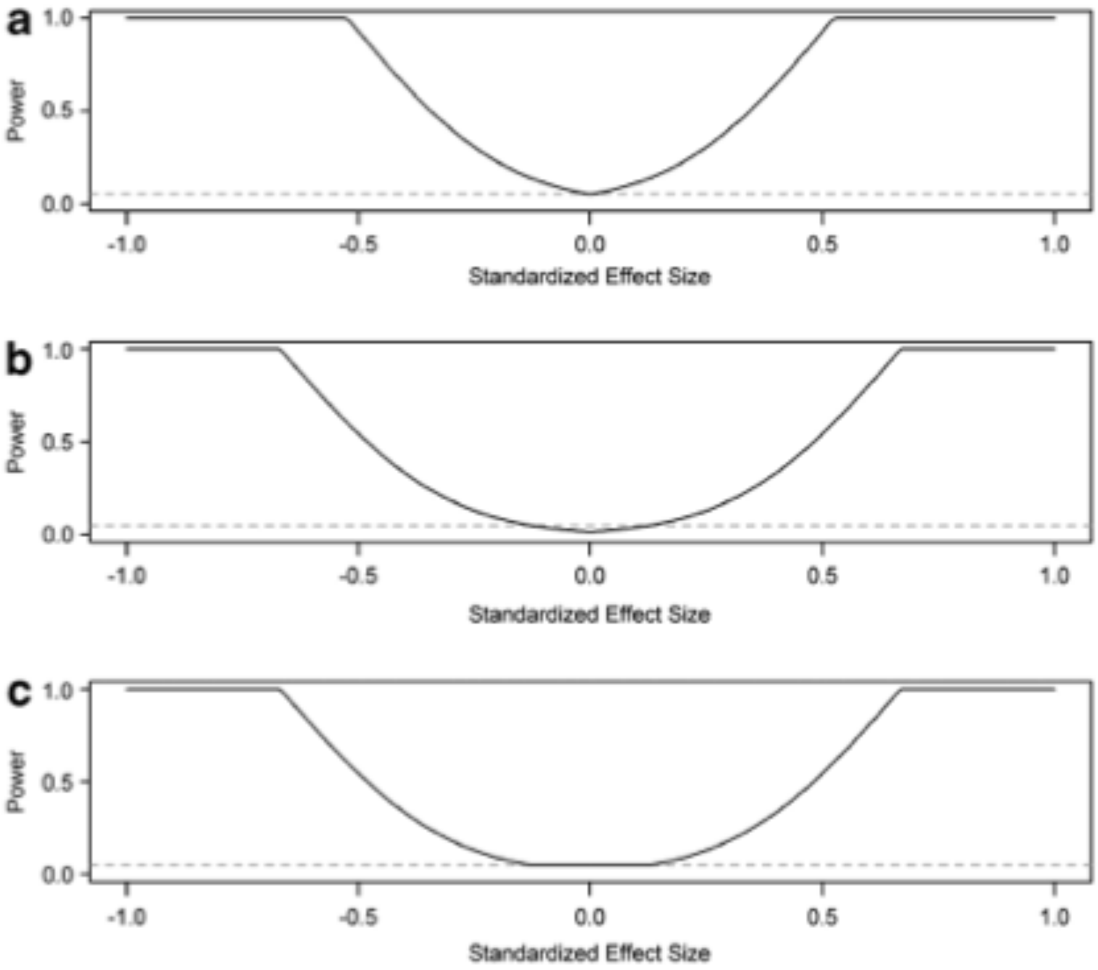
$$\begin{matrix} \dfrac{|\hat{\beta}|-\beta_0}{\mathrm{SE}(\hat{\beta})} & \text{if } |\hat{\beta}| \geq \beta_0 \\ 0 & \text{otherwise} \end{matrix}$$

if your effect size is smaller than your measurement error, it’s not interesting. ppl argue that the m.e. is small, but the effect is still smaller.
if you cannot make more accurate measurements, then your study is not interesting

Where β_0 corresponds to the limit of measurement precision, not zero.

European Journal of Human Genetics (2014) 22, 158–159; doi:10.1038/ejhg.2013.110; published online 22 May 2013

Protected Inference



European Journal of Human Genetics (2014) 22, 158–159; doi:10.1038/ejhg.2013.110; published online 22 May 2013

R^2: The co-efficient of determination

- R^2 is probably the most commonly used quantity for model fit.

- Often described as the proportion of variation *explained* by the model.

unless you're testing for a causal relationship, you're not explaining anything. use accounted

- I prefer: proportion of variation *accounted* for by the model.

- I really prefer thinking about 1 -R^2: proportion of variation for unaccounted for (how much are you missing the mark).

R^2 is not good for quantifying model fit

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).

cannot compare R and R^2. R has directionality, R^2 is association between two variables.

unadjusted: the more parameters you add, R^2 will increase. we need to adjust for this. if you add parameters with no information, adjusted R^2 will decrease

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.

Generalized R^2

$$\text{Generalized R}^2 = \frac{1 - \left(\dfrac{L(\text{null})}{L(\hat{\theta})} \right)^{2/n}}{1 - L(\text{null})^{2/n}}$$

Where L(null) is the log Likelihood for the null model, L(theta hat) is the log likelihood for the MLE of the model, and n is sample size.

This does not adjust for parameters.

Is R^2 useful?

- Yes. It is very useful in making a statement about **overall** model fit (% variation accounted for).
- But, it is ***not useful*** in the 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?

Model vs predictor specific R^2

- Can we just fit individual models for each predictor to calculate the R^2?

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)
```

Coefficient of Partial Determination Partial R^2

The partial R^2 for X$_1$, given that X$_2$ is already in the model is calculated as

$$R^2_{Y1|2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSM(X_1 | X_2)}{SSE(X_2)}$$

portion of variation accounted for by each individual variable (including interactions)
? not sure, follow up

This extends more generally.

$$R^2_{Y4|123} = \frac{SSM(X_4 | X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}$$

SSE = Error (residual) Sum of Squares
SSM = Model (regression) sum of squares (fitted)

Salient Points

- There are several classes of effect sizes (unstandardized, scaled by pooled sd, scaled by mean, variance accounted for, odds ratios).
- Deciding which to use depends on the question at hand, and with what you compare your results to.
- This can take a considerable amount of thought.

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!