Background: Effect Size

$$\text{Effect size} = \frac{\text{[Mean of experimental group]} - \text{[Mean of control group]}}{\text{standard deviation}}$$

- Expressed in relevant units
- Not just “significant” – *how* significant?
- Used prolifically in meta-analysis to combine results from multiple studies
  - But be careful – averaging results from different experiments can produce nonsense

Caveat: Other definitions of effect size exist: odds-ratio, correlation coefficient

Robert Coe, 2002, Annual Conference of the British Educational Research Association
It's the Effect Size, Stupid: What effect size is and why it is important.

Effect Size

- Standardized Mean Difference

$$ES = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{pooled}}$$

Lots of ways to estimate the pooled standard deviation

$$\sigma_{pooled} = \hat{\sigma}_2$$

Glass, 1976

$$\sigma_{pooled} = \sqrt{\frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}}$$

e.g., Hartung et al., 2008

Meta-analysis

- 1978: Gene V. Glass statistically aggregate the findings of 375 psychotherapy outcome studies Glass (and colleague Smith) to disprove claim that psychotherapy was useless
- Glass coined the term “meta-analysis”
- Earlier ideas from Fisher (1944)
  - “When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance”

adapted from slide by David B. Wilson, 1999

Meta-analysis

- Even more important in data science
  - You will often be working with data you didn't collect
  - "Big Data" may have become big by combining data from different sources
  - When is this ok? Test for homogeneity

Meta-analysis: Weighted Average

- Idea: Average across multiple studies, but give more weight to more precise studies
- Simple method: Weight by sample size

$$w_i = \frac{n_i}{\sum_j n_j}$$

- Inverse-variance weight =

$$w_i = \frac{1}{se^2}$$

Lots of variants

Caveat: This is a fixed-effect model: it assumes that each individual study is measuring the same true effect. We won't talk about the random effects model.

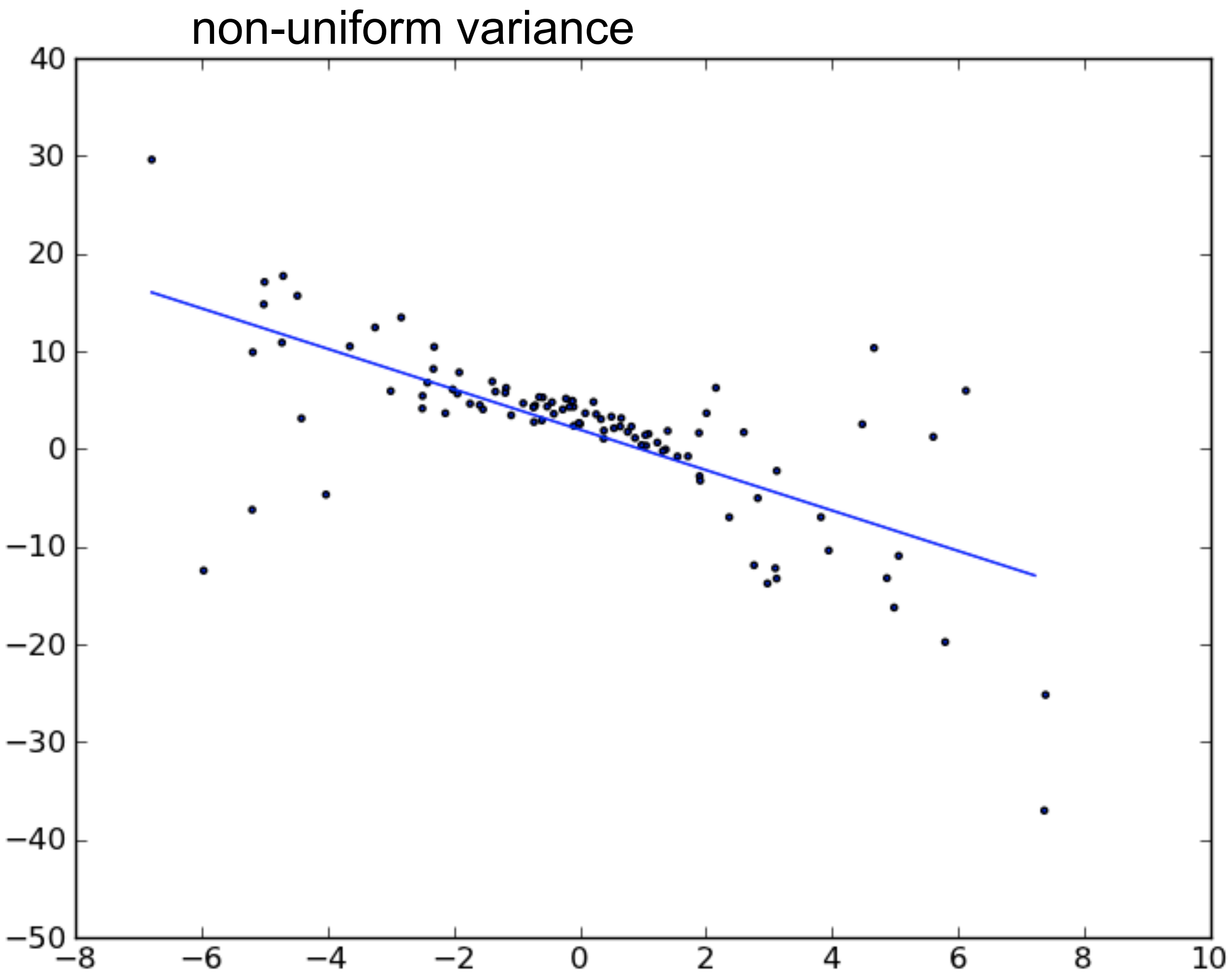
Effect size: Cohen's Heuristic

- Standardized mean difference effect size
  - small = 0.20
  - medium = 0.50
  - large = 0.80

Confidence Interval (of effect size)

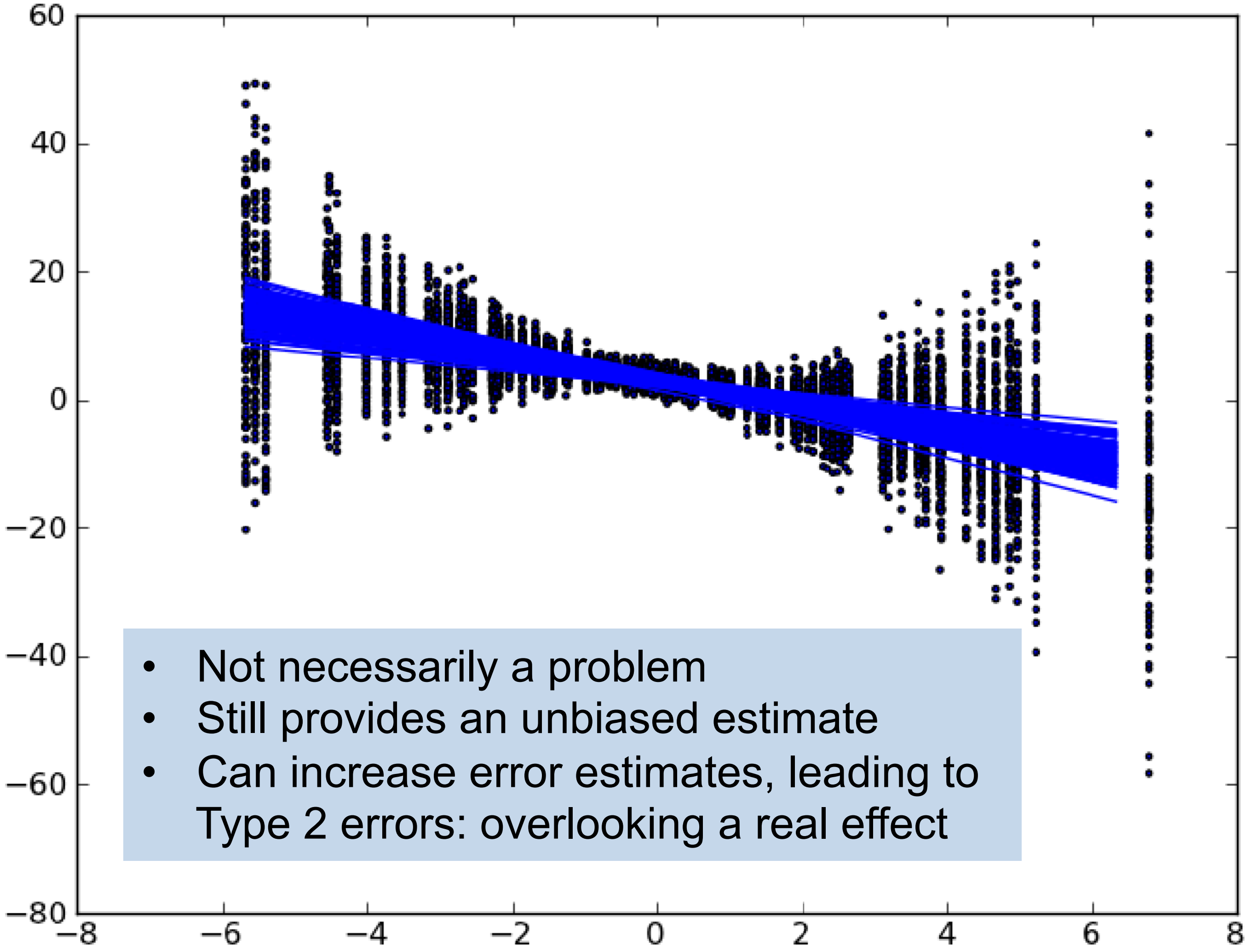
- What does a 95% confidence interval of the effect size mean?
  - If we repeated the experiment 100 times, we expect that the interval would include this effect size 95/100 times
  - If this interval includes 0.0, that's equivalent to saying the result is not statistically significant.

Aside: Heteroskedasticity



Aside: Heteroskedasticity

100 repetitions, same x-values, y-values drawn from the same model



- Not necessarily a problem
- Still provides an unbiased estimate
- Can increase error estimates, leading to Type 2 errors: overlooking a real effect