Thinking Statistically

What Counts and What Doesn't?

Acknowledgements

This work is partly supported by the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. "IDEAL"

My thanks to CASI for further support and Gabrielle Kelly and the oganisers for the invitation

Four ‘paradoxes’

- Covariates measured with error in randomised clinical trials
  - Easy
- Meta-analysis of sequential trials
  - Very hard
- Dawid’s selection paradox
  - Fairly hard
- Publication bias in the medical literature
  - Both hard and easy

To get you thinking

- A series of independent binary events with common probability θ of a 'success'
- You have a uniform prior for θ, $f(\theta) = 1, 0 \leq \theta \leq 1$
- You will carry out an experiment in which one million trials will be performed
- Which is more likely:
  - You will witness exactly one million successes
  - You will witness 500,000 successes and 500,000 failures in any order?

Covariate errors

Base logic

Regression dilution bias

The well known estimate of the slope for regression equation for *Y* on *X* is

$$\hat{\beta}_{Y|X} = \frac{cov(XY)}{var(X)}$$

Now suppose that you measure *X* with a certain amount of random error to obtain *x.* The regression of *Y* on *x* is now

$$\hat{\beta}_{Y|x} = \frac{cov(xY)}{var(x)}$$

However, although on a plausible model $E[cov(xY)] = E[cov(XY)]$ we shall find that in general

$$var(x) > var(X)$$

Hence it follows that

$$\hat{\beta}_{Y|x} < \hat{\beta}_{Y|X}.$$

This is referred to as *regression dilution bias*

Many have concluded that in consequence ANCOVA is not conditionally unbiased

The argument

- We observe a difference in a covariate of Δ between the groups
- We need to adjust for Δ to have a conditionally unbiased estimate
- Due to regression dilution bias we under-adjust
- The resulting estimate is biased

The proof

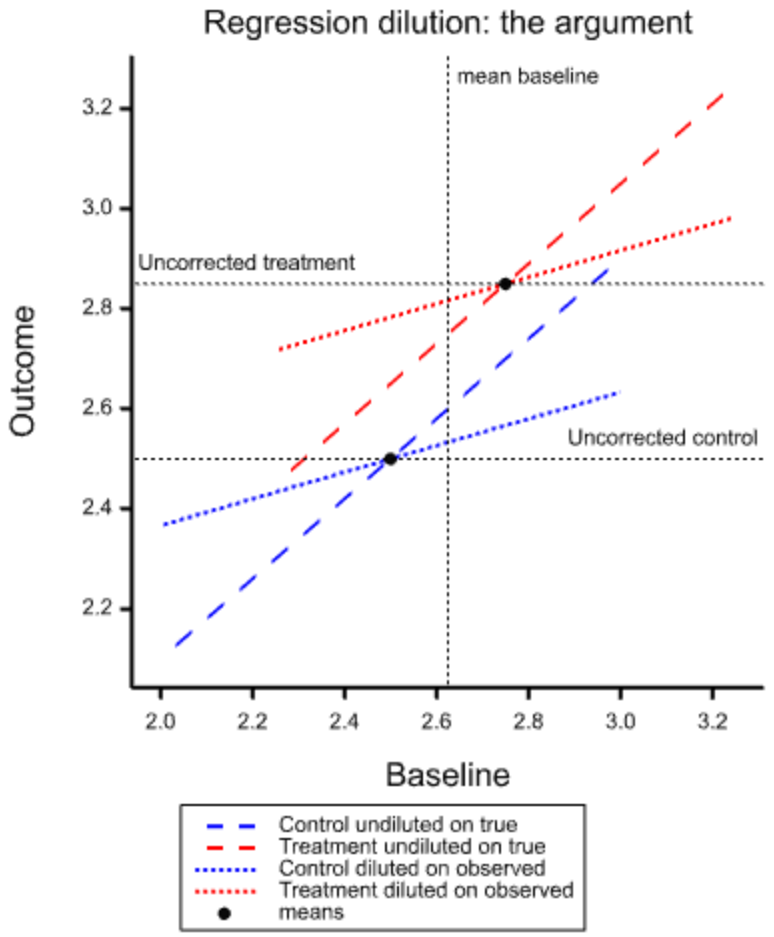
- Take a bivariate Normal
- Consider a case where the true difference in the covariate is Δ
- Work through the expectation
- Show the estimate is biased
- Simulate also to demonstrate it

Graphical ‘demonstration’

The regression lines for outcome on observed baselines are shallower than for outcomes on true baselines

Hence adjustment is less

Thus, it is claimed, the treatment estimate is conditionally biased

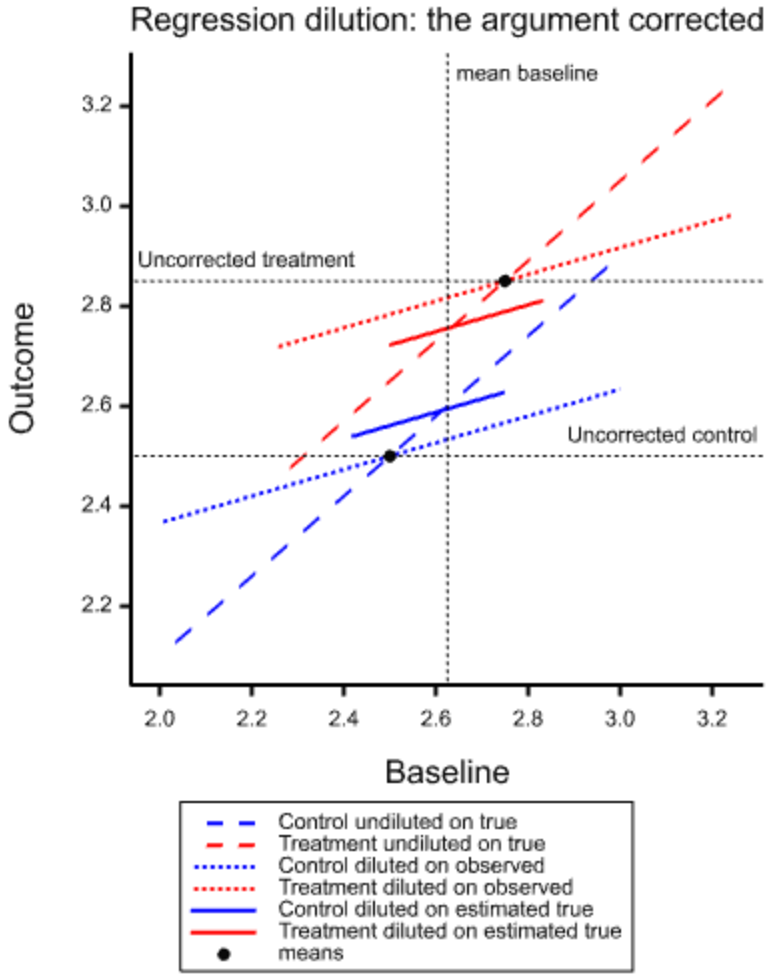


The argument is false

It overlooks the fact that since the true values will have a lower variance than the observed ones we must expect that if they were observed the groups would be closer to each other (and also less diffuse)

When this further effect is accounted for ANCOVA is seen to be conditionally unbiased

However, the whole argument is unnecessary if one thinks clearly in the first case



The difference between the true regression lines and the diluted regression lines is greater if plotted against the observed values but is identical if plotted against the *expected true* values.

(The distance between thee three pairs of parallel lines can be judged by comparing them at the overall mean.)

Thinking clearly

- If you have two covariates measured without error you adjust to the extent that they are predictive
  - If one predicts more than another you adjust more
  - It's the degree to which something is predictive that governs the adjustment
- If you don't measure a covariate you can't condition on it
  - But ANCOVA is still unbiased because you have an RCT
- So you can't do better than condition on what you observe
- Trivially, the observed covariate is what you have observed
  - You adjust for what you have observed to the extent that what you have observed is predictive
- You haven't observed the true covariate
  - Its regression is irrelevant
  - Because it is unobserved
- ANCOVA is conditionally unbiased conditioning on the observed covariate and that is all that matters

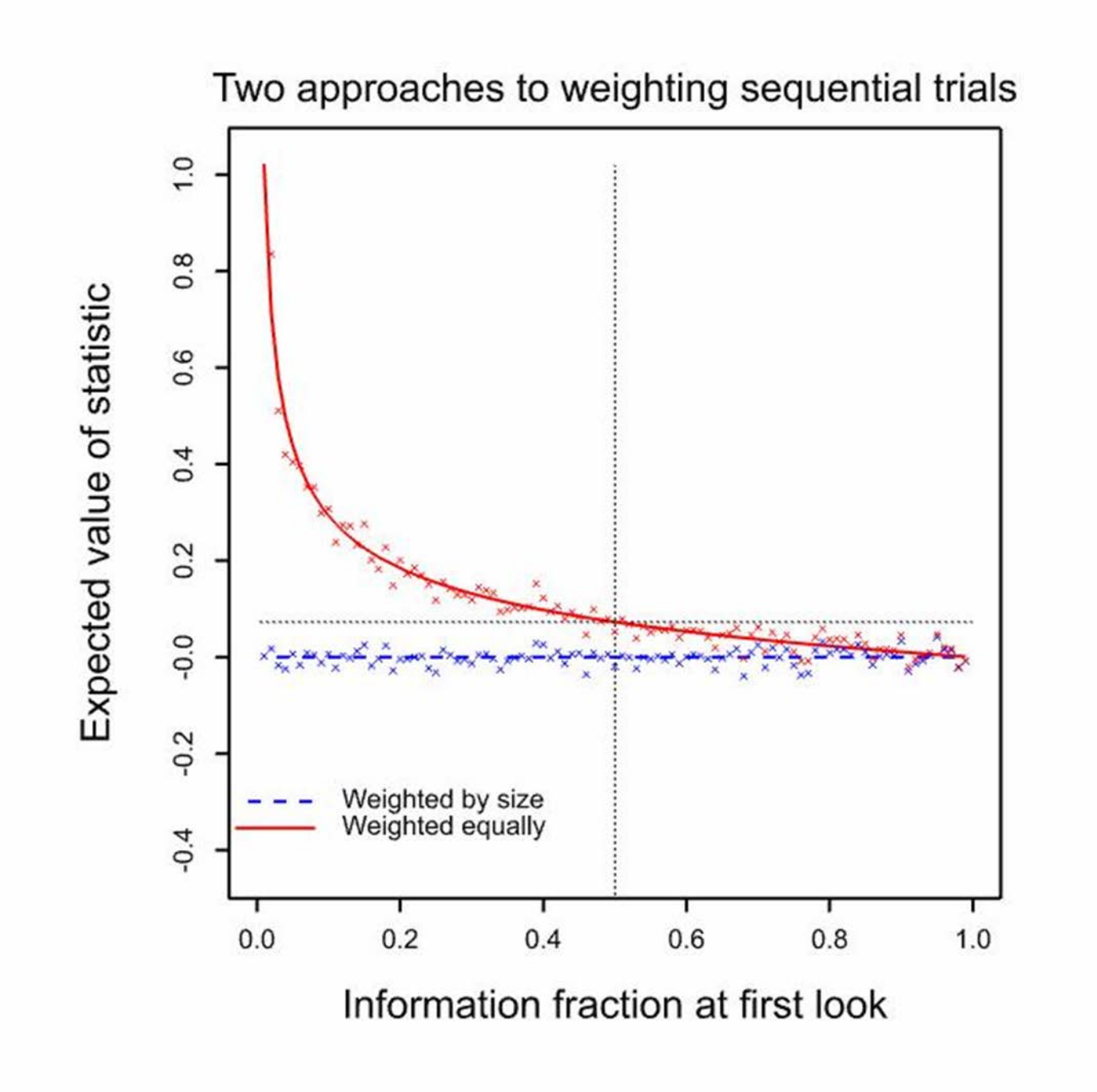
Sequential meta-analysis

Weight to go

The problem

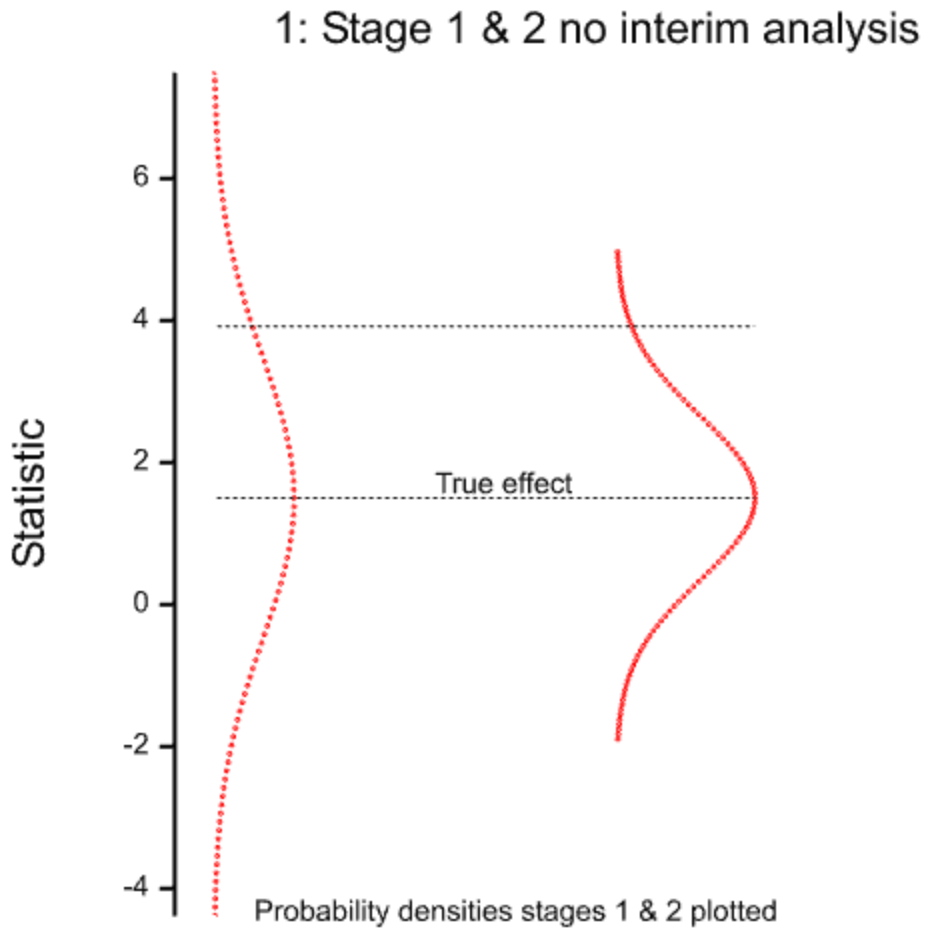
- It is well known that any good frequentist should pay a penalty for monitoring a clinical trial with the intention to stop for efficacy
  - All drug regulators agencies will require this
- Suppose that a number of such trials have been run
- It is proposed to perform a meta-analysis
  - A formal summary of the results
- Is an adjustment needed for stopping?

Answer:
“no”


Two approaches to weighting sequential trials
Expected value of statistic
1.0
0.8
0.6
0.4
0.2
-0.0
-0.2
-0.4
Weighted by size
Weighted equally
0.0
0.2
0.4
0.6
0.8
1.0
Information fraction at first look

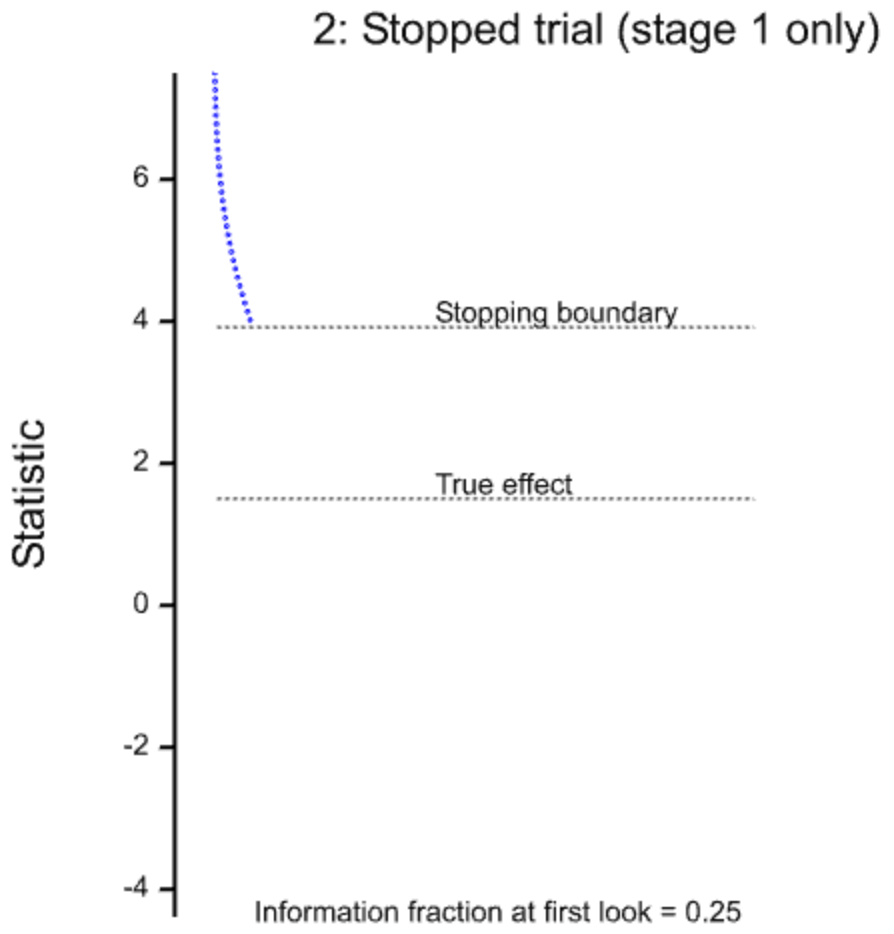
It is assumed that you look after an information fraction of 1/4

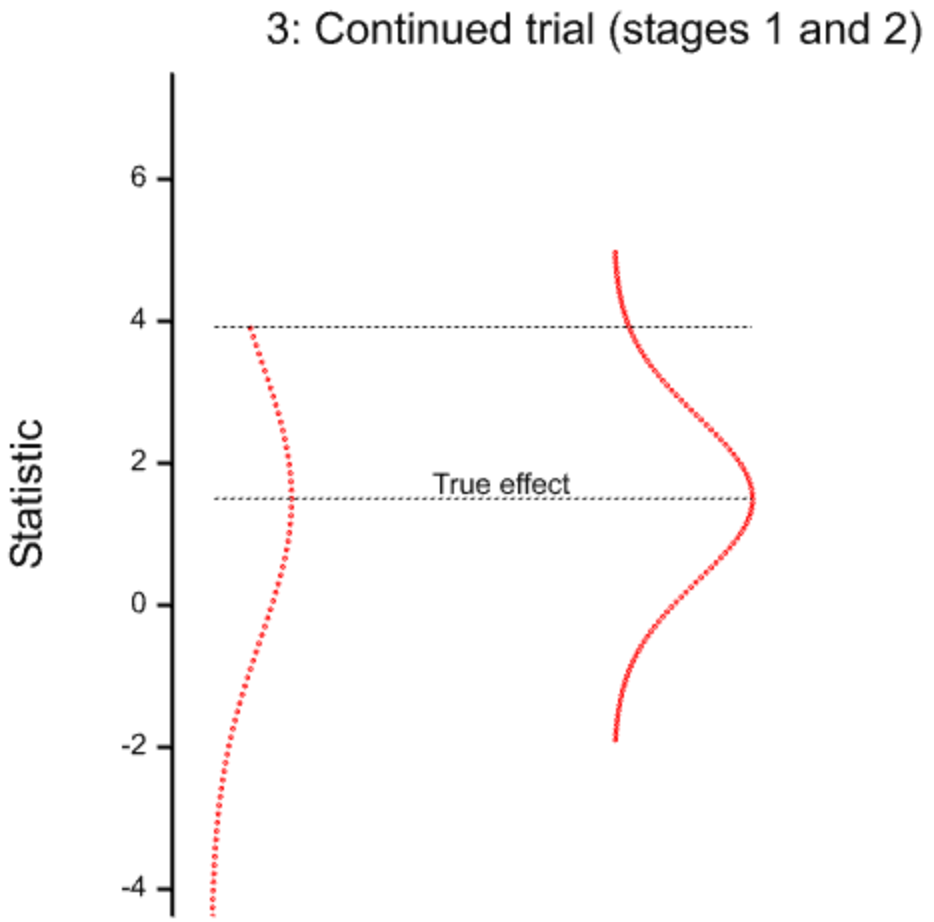
These are (expected) probability densities of the estimated treatment effects

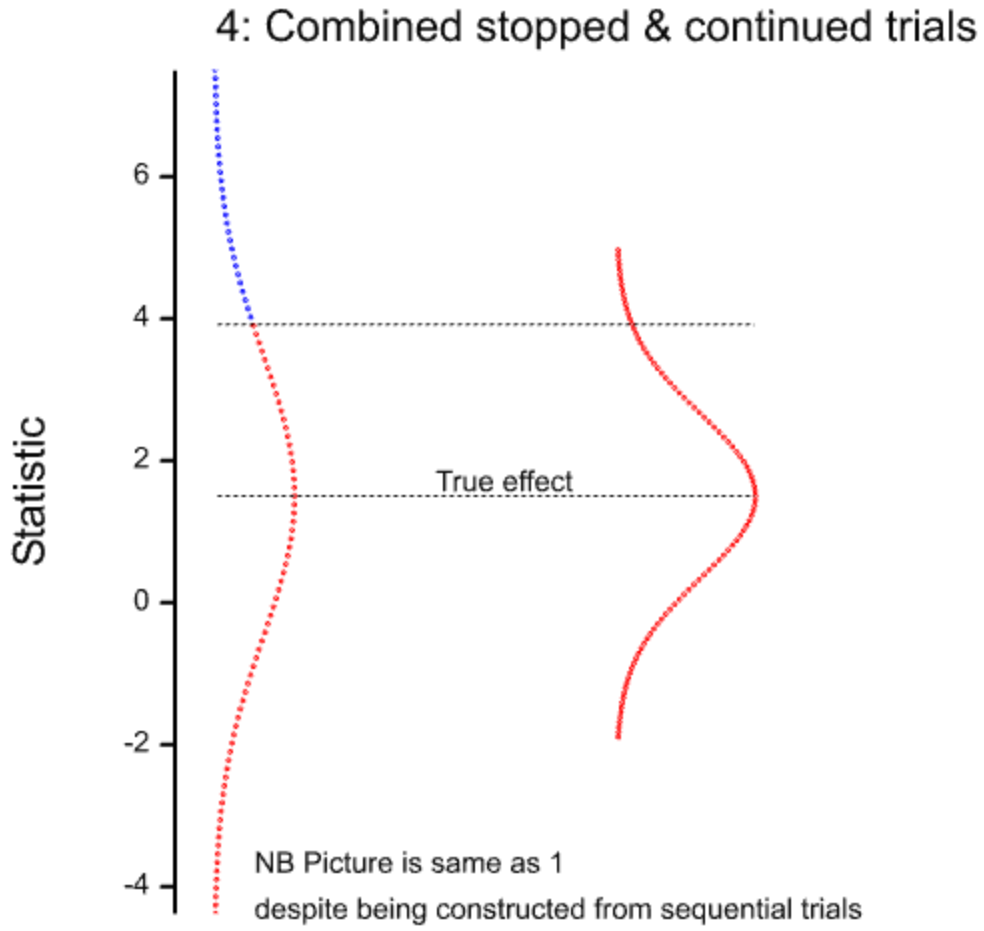


The right hand curve is what applies later in the trial and since it is based on more information is narrower

NB this is not the density; it does not integrate to 1. It is the density multiplied by the probability of the trial stopping early

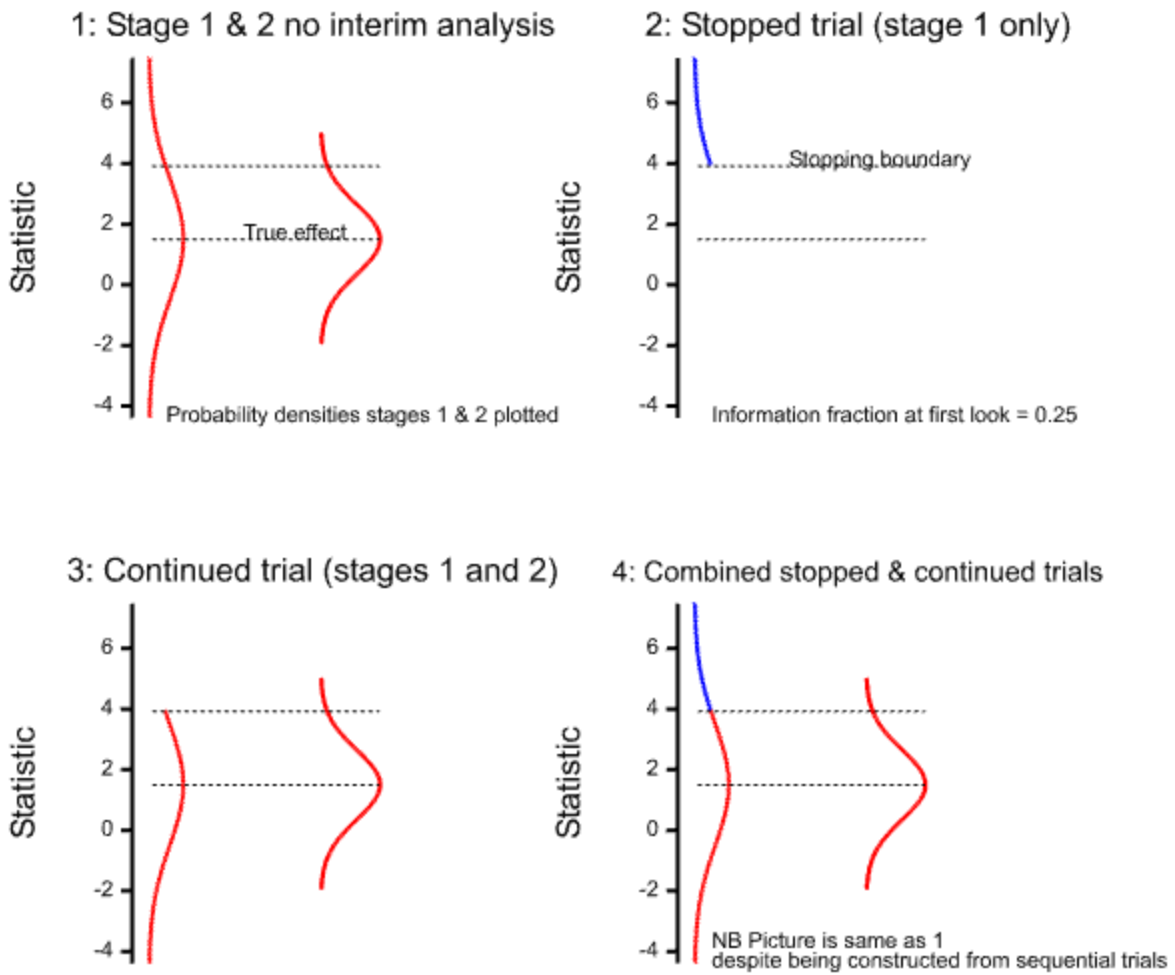



3: Continued trial (stages 1 and 2)
Statistic
6
4
2
0
-2
-4
True effect


4: Combined stopped & continued trials
Statistic
6
4
2
0
-2
-4
True effect
NB Picture is same as 1
despite being constructed from sequential trials

Conclusion

- You can put the trials together in a way that gives an unbiased estimate
- You need to weight the results by the amount of information
- Trials that stop early will get less weight
  - They have the capacity to overestimate treatment benefit
- Trials that continue will get more weight
  - They have the capacity to underestimate treatment benefit
- Weighting by information deals with the problem
- This is what fixed effects meta-analysis does


1: Stage 1 & 2 no interim analysis
Statistic
6
4
2
0
-2
-4
True.effect
Probability densities stages 1 & 2 plotted
2: Stopped trial (stage 1 only)
Statistic
Stopping.boundary
Information fraction at first look = 0.25
3: Continued trial (stages 1 and 2)
Statistic
4: Combined stopped & continued trials
Statistic
NB Picture is same as 1
despite being constructed from sequential trials

Dawid's selection paradox

Choices, choices

The paradox

Philip Dawid, in a paper of 1994, drew attention to a clash between a frequentist intuition that the interpretation of a data set ought to be different if the data were specifically chosen for some feature

Example the best of a number of treatments being studied

However, this does not (usually) matter for the Bayesian

“Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data.”

A Selection Paradox of Dawid's

- Suppose that we estimate treatment means from a number of treatments in clinical research
- We use a standard conjugate prior
- Since Bayesian analysis is full conditioned on the data, then for any treatment the posterior mean will ***not*** depend on why we have chosen the treatment
  - At random
  - Because it gave the largest response

See DAWID, A. P. (1994), in *Multivariate Analysis and its Applications*, eds. T. W. Anderson, K. a.-t. a. Fang, & I. Olkin

A Model

$$X_i \sim N\left(\mu_i, \sigma^2\right)$$ Observed mean for treatment *i*

$$\mu_i \sim N\left(\theta, \tau^2\right)$$ Prior distribution (assumed iid for all treatments)

$$\theta, \tau$$ Known prior parameter

Since θ, τ are known, without loss of generality we can measure everything in terms of the standardised units of the prior distribution.

We assume in what follows this has been done already, so that $\theta = 0$, $\tau = 1$ and the other symbols can remain unchanged. Thus σ^2 becomes the ratio of data variance to prior variance.

Posterior inference

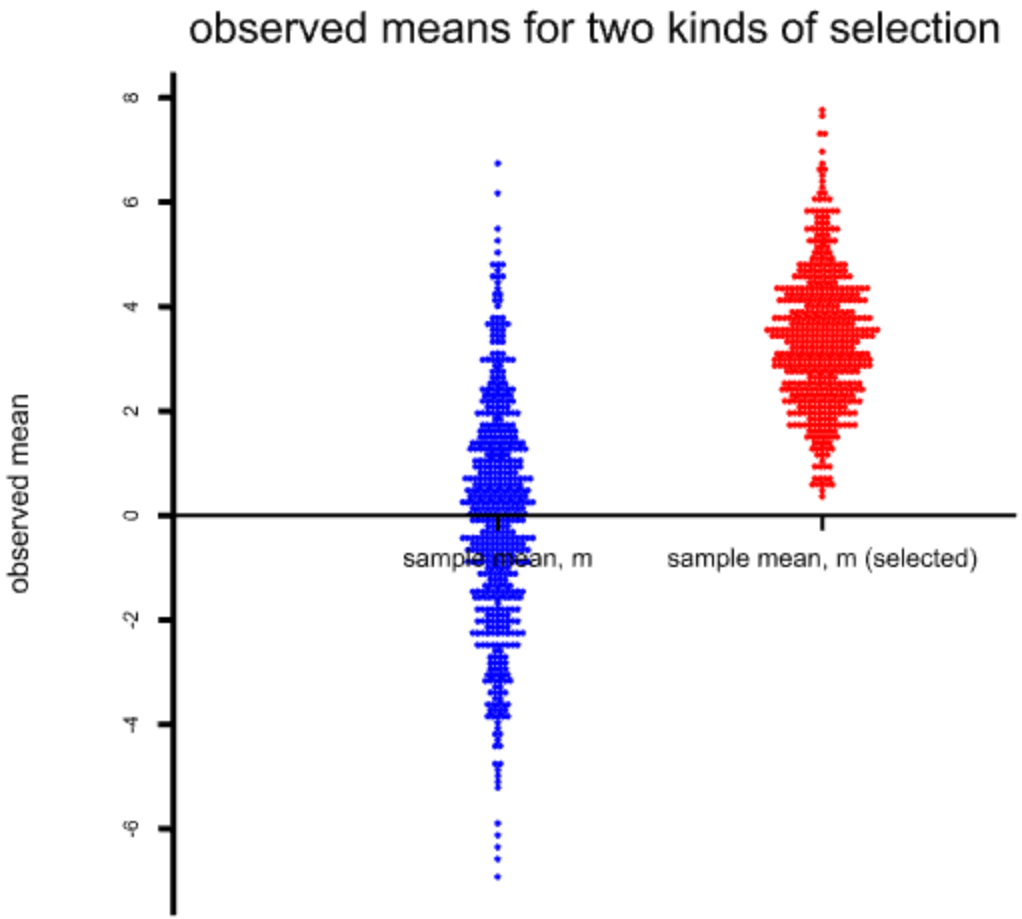
Let y_i be the posterior mean corresponding to data mean x_i and let q_i be the posterior variance. Hence, we have from standard Bayesian results, whereby the posterior mean is the precision-weighted linear combination of prior and data means and the posterior precision is the sum of prior and data precisions ('precision' here being the reciprocal of the variance) (Box and Tiao, 1992),

$$y_i = \left(\frac{1}{1+\sigma^2}\right)x_i, \quad q_i^2 = \left(1+\frac{1}{\sigma^2}\right)^{-1}, \quad \mu_i \sim N\left(y_i, q_i^2\right).$$

If x^* is the largest data mean, y^* the corresponding posterior mean and μ^* the corresponding "true" mean, then all we need to do to obtain the corresponding inference is to substitute x^* for x_i in the above.

NB the frequentist estimates are just the observed mean.

These are indeed higher if the means were selected as highest values in a set.

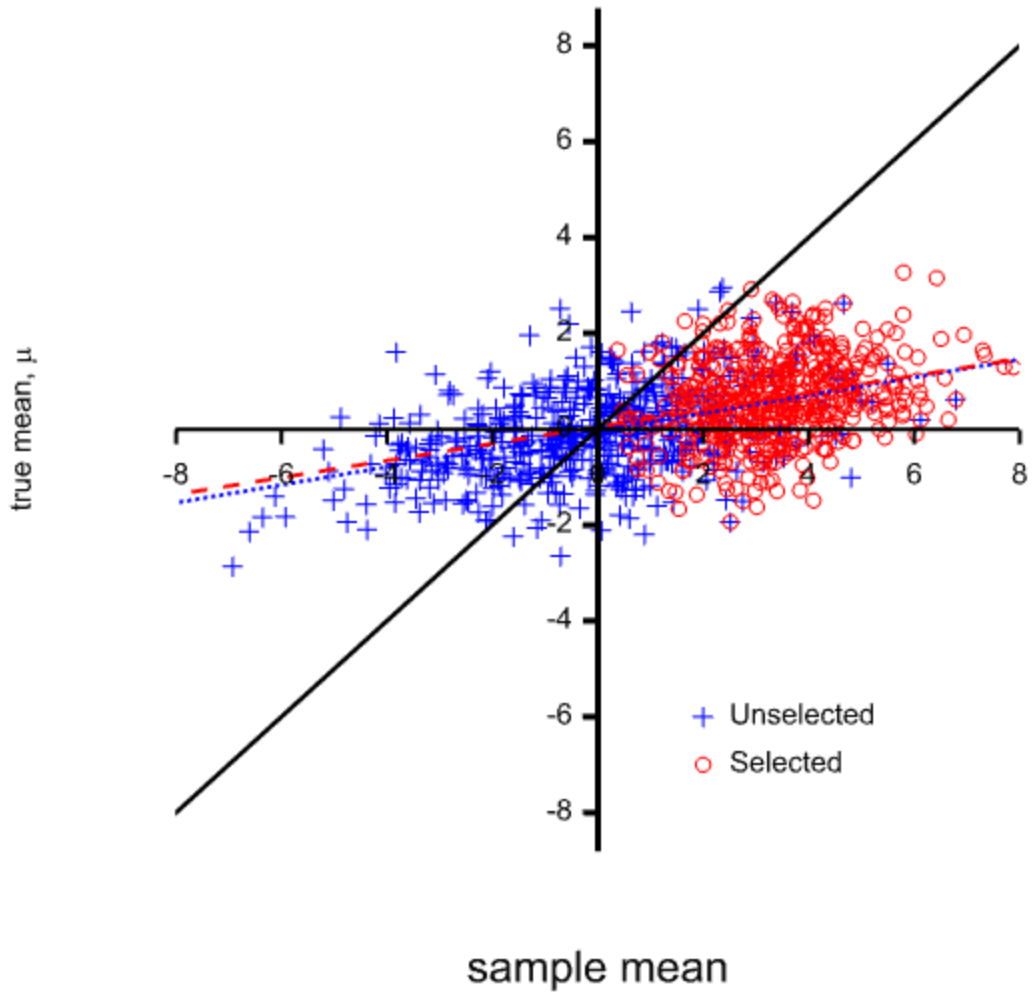


NB the Bayesian selection is given by the theoretical regression.

This example is a simulation from the corresponding prior

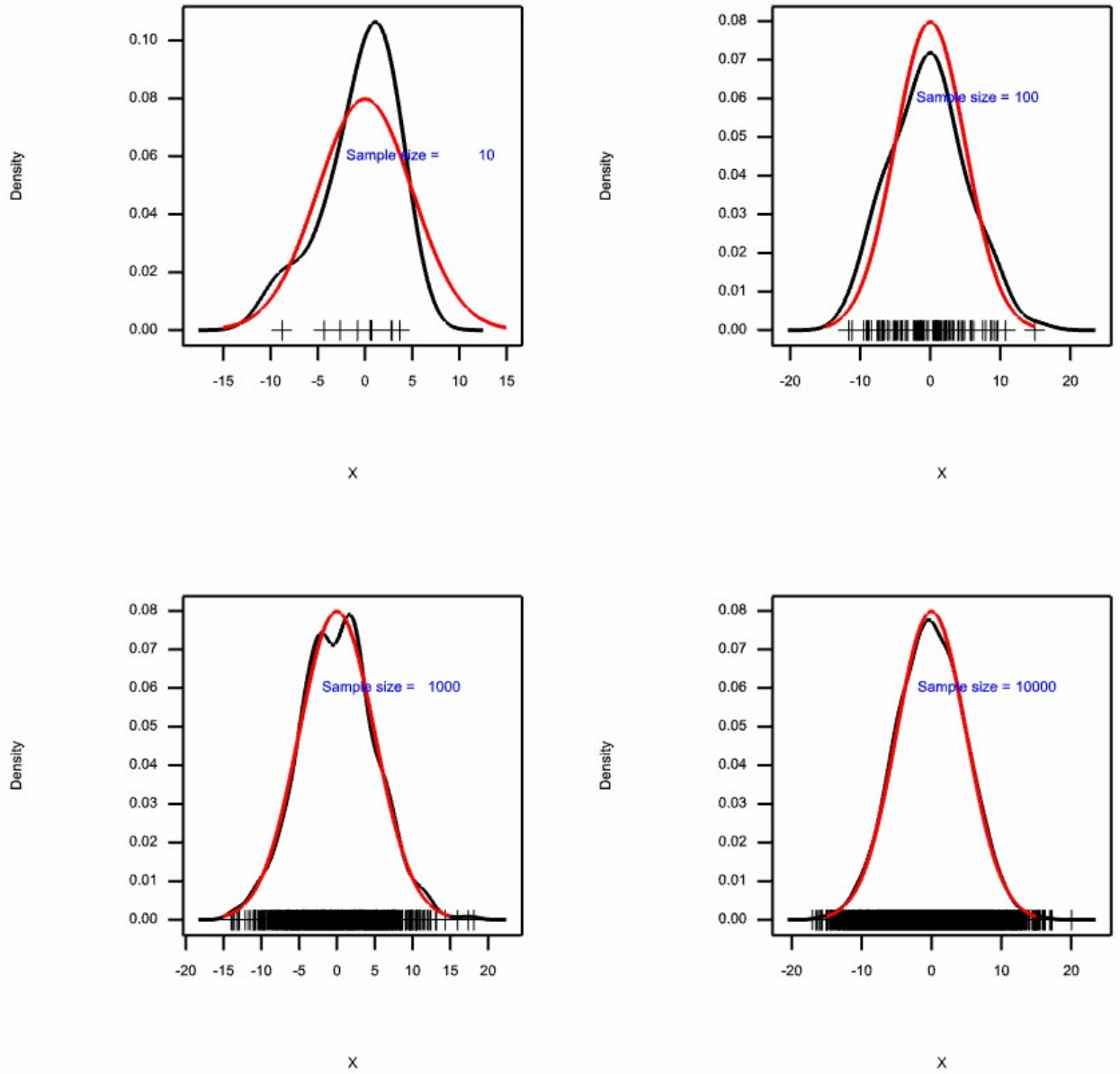
The regression is the same for the two cases.

Thus given the mean it makes no difference to the inference whether it was selected or not



A Heuristic Explanation

- A Normal prior distribution for the true means could be expressed in two ways
  - Parametrically using two parameters
  - Non-parametrically using an empirical distribution
- The more means are represented in the empirical distribution the closer it comes to the parametric representation
- Hence, having a parametric prior is equivalent to having observed an infinity of true means
  - You are 100% certain about the mean of your prior


Density
0.10
0.08
0.06
0.04
0.02
0.00
Sample size = 10
-15 -10 -5 0 5 10 15
X
Density
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Sample size = 100
-20 -10 0 10 20
X
Density
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Sample size = 1000
-20 -15 -10 -5 0 5 10 15 20
X
Density
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Sample size = 10000
-20 -10 0 10 20
X

Heuristics continued

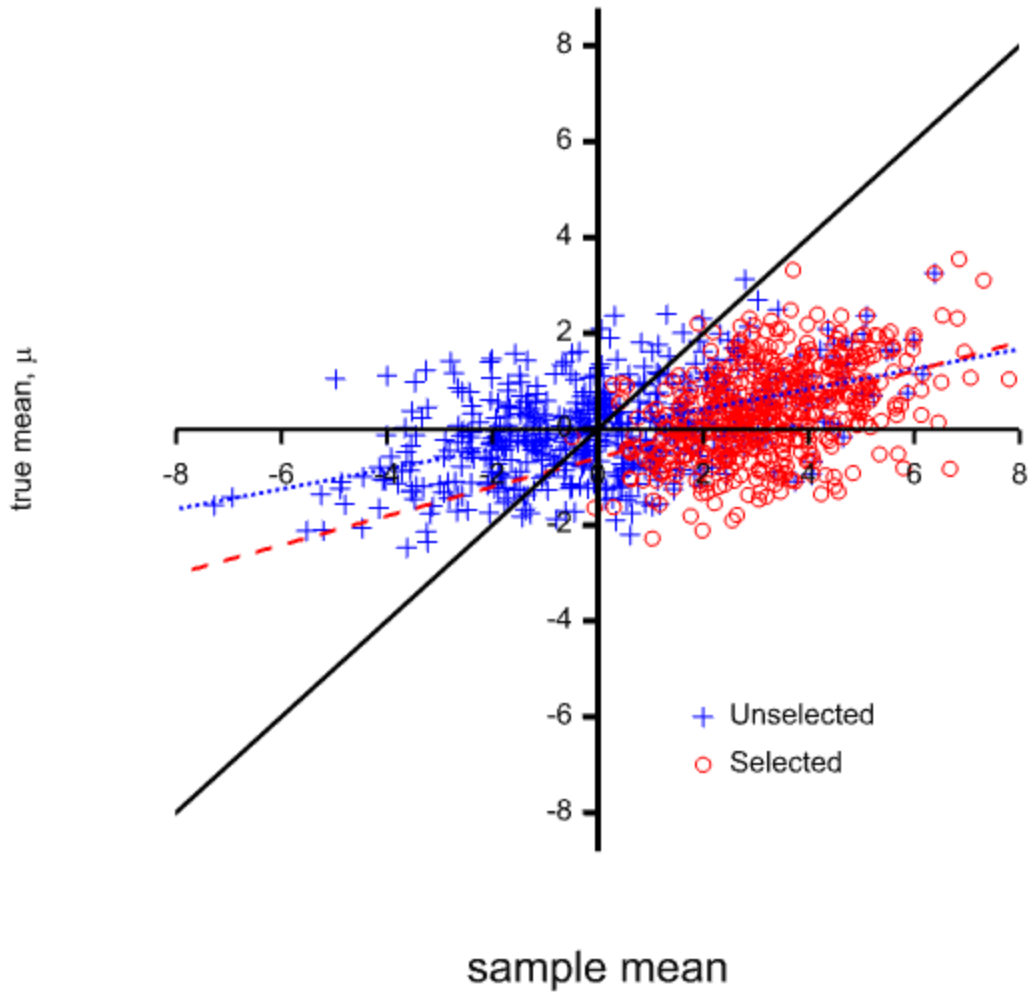
- Hence you know the relative frequency (density) for any one of these infinite true means
- You just don't know which of these true means belongs with your observed one
- Given this infinity of background knowledge the fact that your mean is the highest amongst a small 'local' group of observed means is irrelevant

This is now the hierarchical case

We have a two stage prior

1) For the treatments within a particular class studied by an experiment
2) For the class within the classes of all treatments

Now the two regression slopes are no longer the same



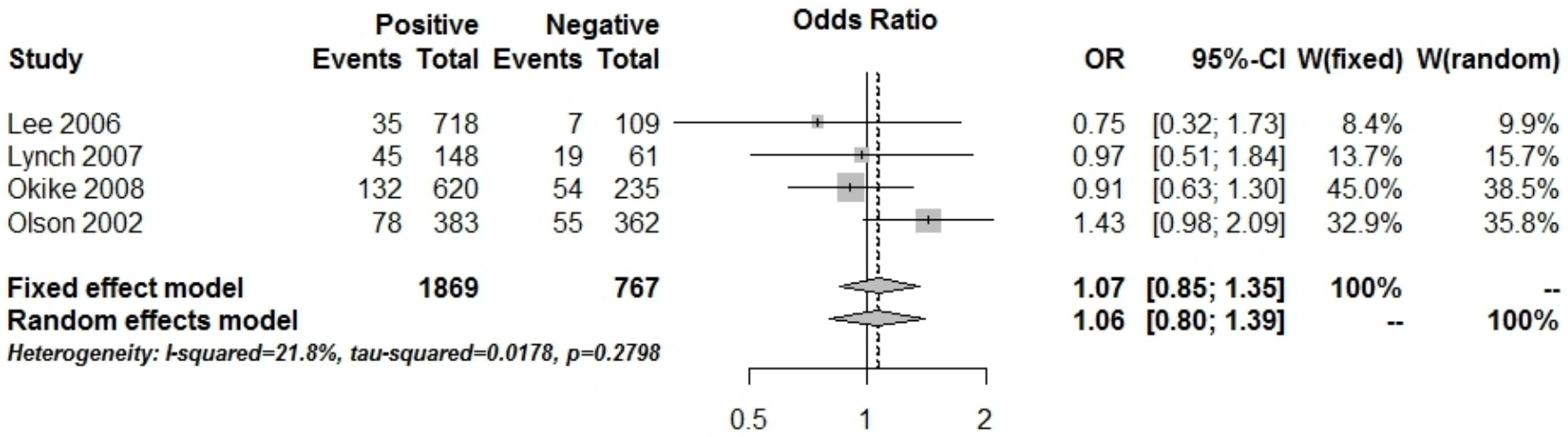
Positive publication bias

Missing inaction

The problem of missing clinical trials

- It seems that positive results are more likely to be published in the literature
- Observational studies in which submissions have been classified as positive or negative seem to show no editorial bias
  - Editors are guiltless
  - Authors are to blame
- A few small experimental studies show the opposite
  - These have been dismissed as unreliable

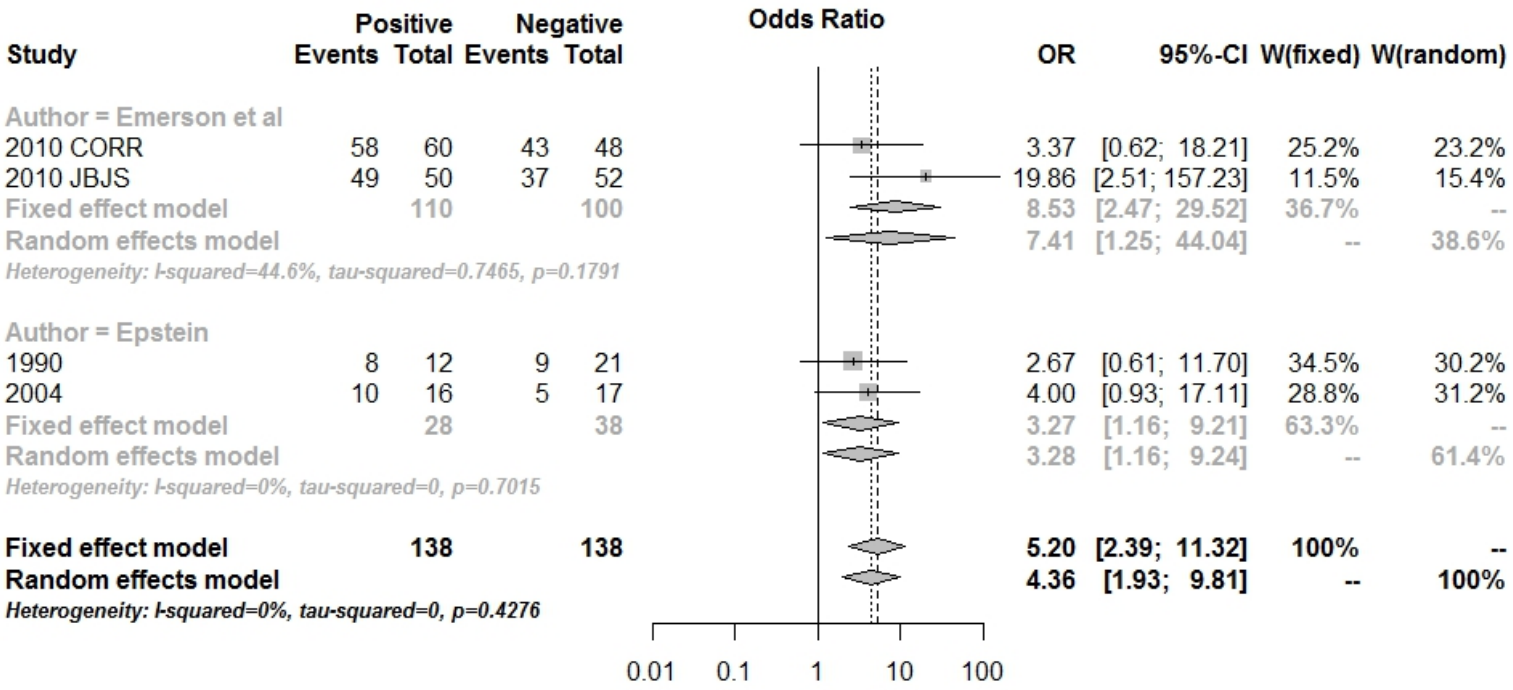
Summary of Observational Studies
(Based on Song et al, 2009)



Favours negative Favours positive

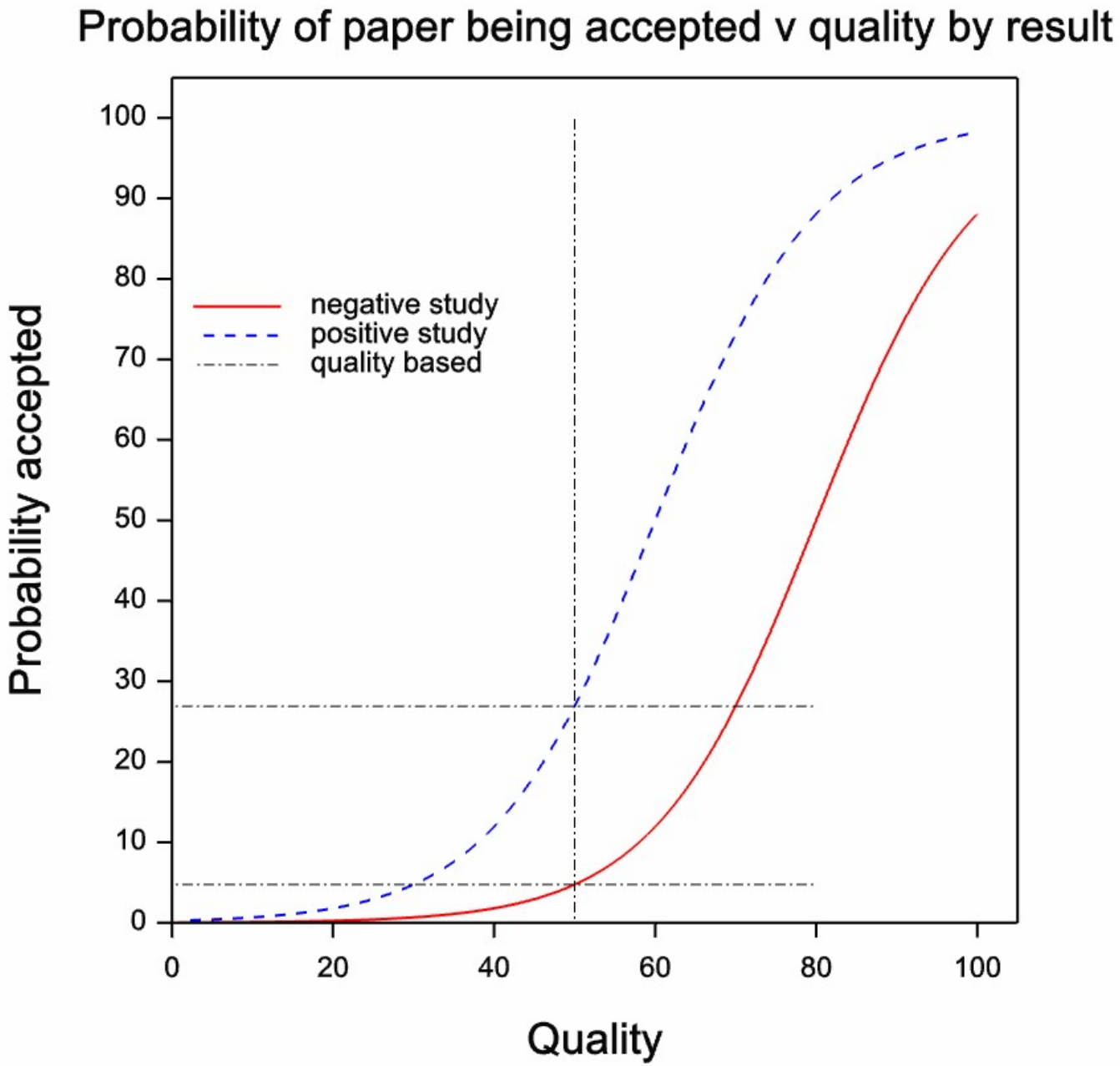
Analysis produced using Guido Schwarzer's *meta* package in R

Summary of randomised studies



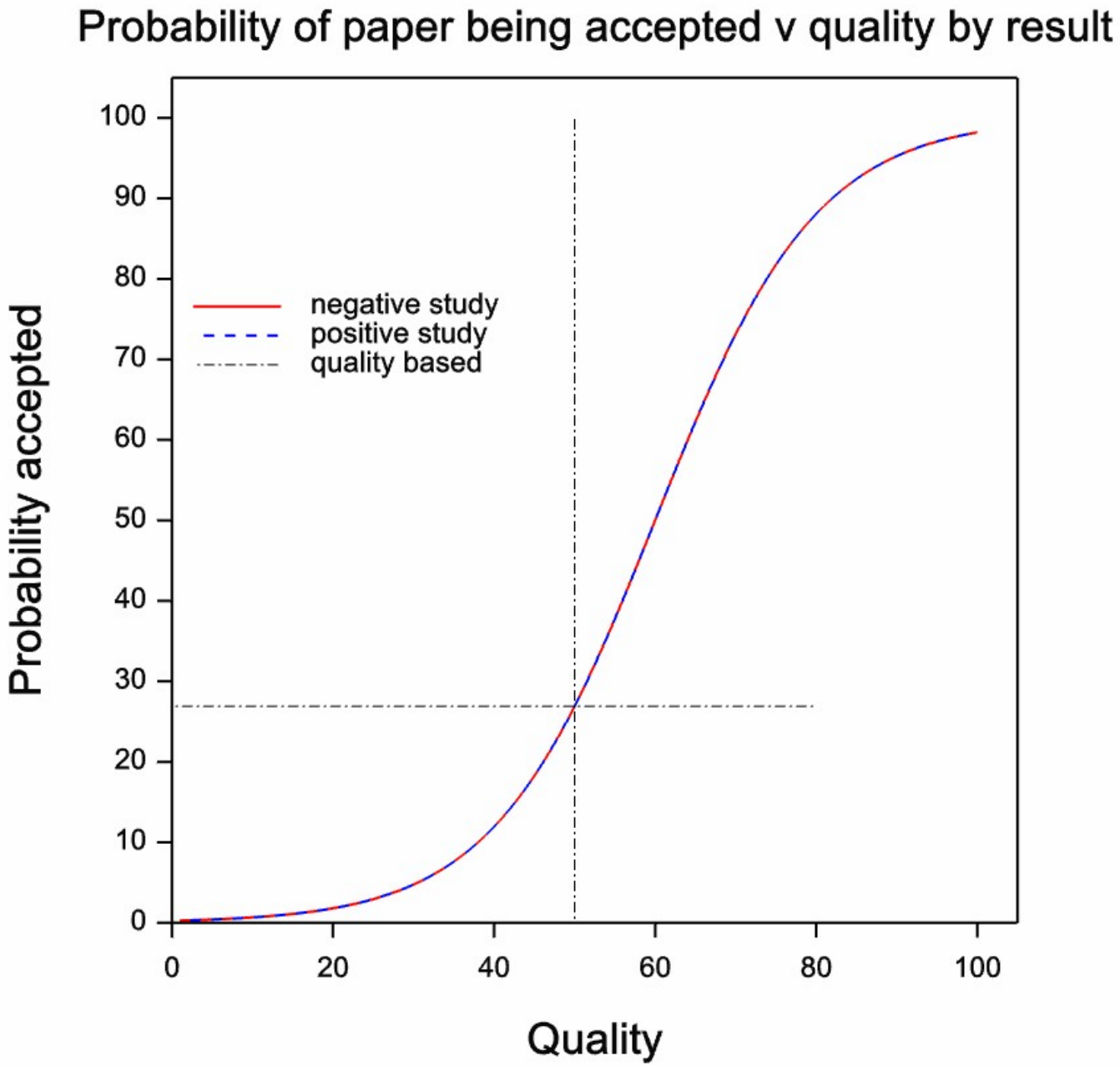
A possible situation describing PQ curves that are not equal

Probability of acceptance increases with quality but is always higher for positive papers



A possible situation describing PQ curves that are equal

Probability of acceptance increases with quality and is identical for both types of paper



Minding your Ps and Qs

- However
- We do not get to see the whole curves
- Not every paper is submitted to every journal
- So the question is what do we see?
- So let's consider two alternatives for curves that differ
  - That is to say for the situation where there is a bias in favour of positive papers

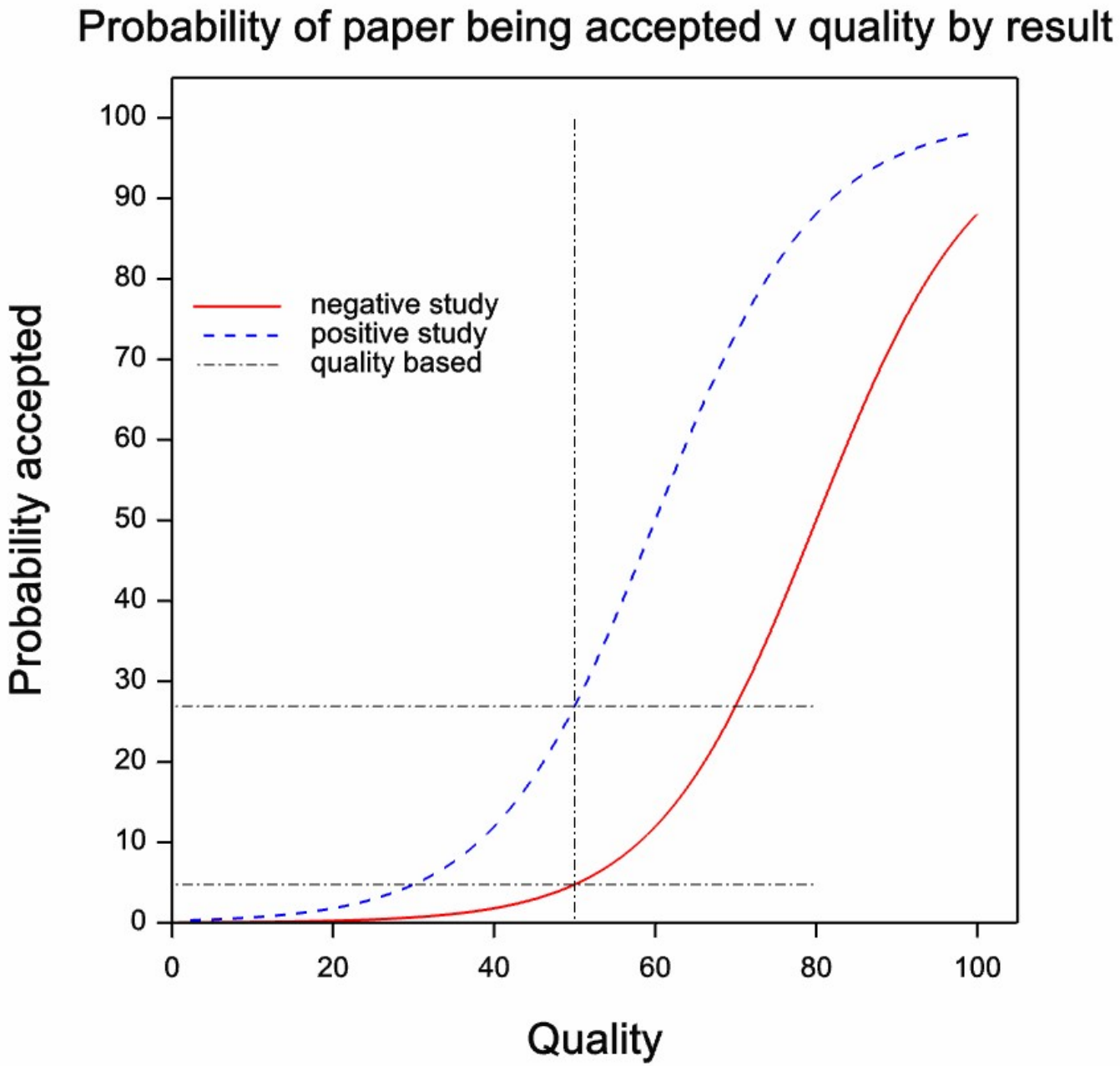
The Q hypothesis
Authors submit papers to journals by quality of the paper

Only quality matters.

In that case if there was a bias in favour of positive papers we would see a difference in acceptance rates.

We don't.

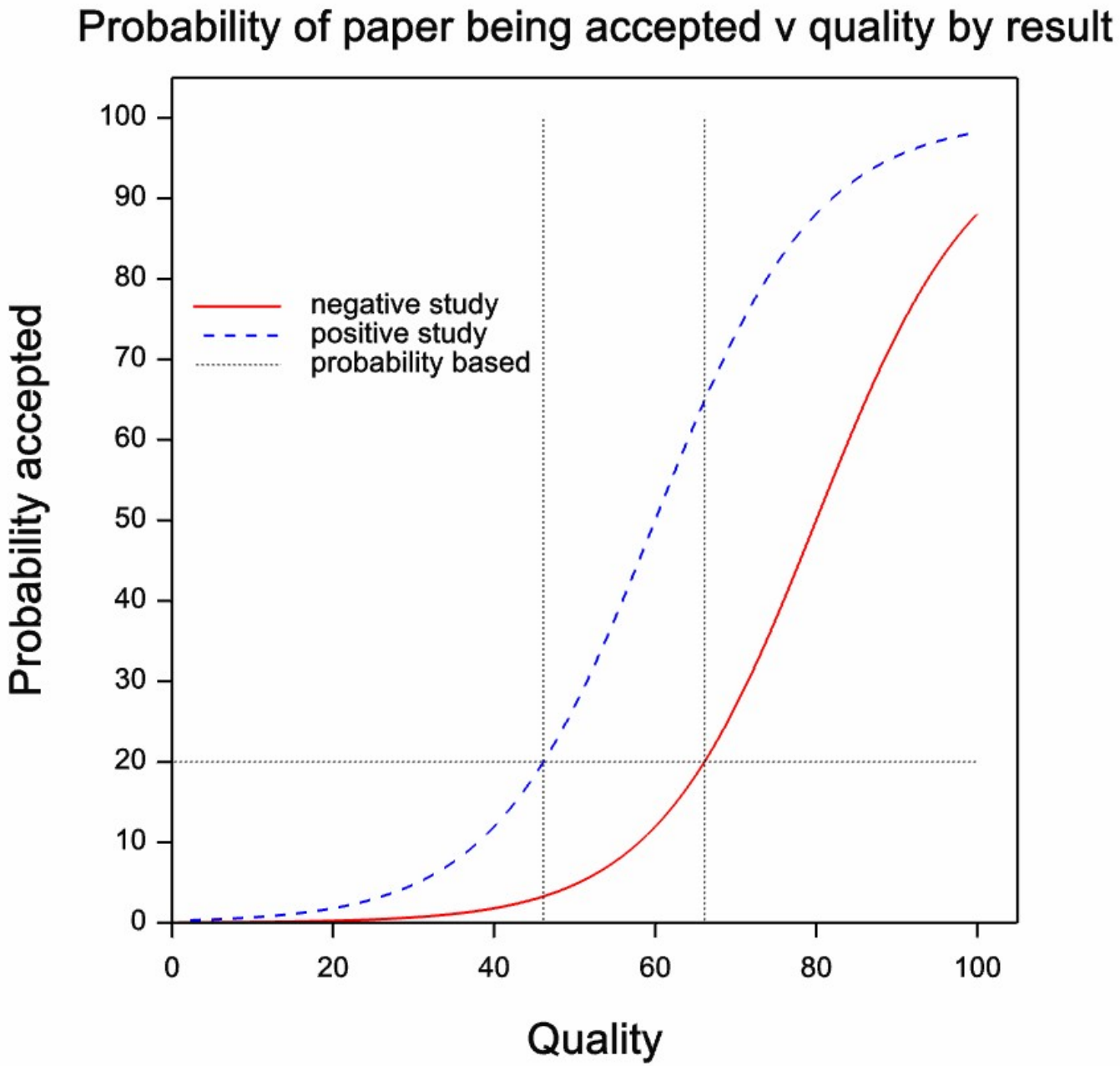
Ergo, there is no bias



The P hypothesis
Authors submit papers to journals by probability of acceptance

To study whether there is a bias or not we need to compare the quality of papers seen in the journal

Is there any evidence it differs?



Commercially Funded and United States-Based Research Is More Likely to Be Published; Good-Quality Studies with Negative Outcomes Are Not

By Joseph R. Lynch, MD, Mary R.A. Cunningham, MD, Winston J. Warme, MD, Douglas C. Schaad, PhD, Fredric M. Wolf, PhD, and Seth S. Leopold, MD

Results: Two hundred and nine manuscripts were reviewed. Commercial funding was not found to be associated with a positive study outcome ($p = 0.668$). Studies with a positive outcome were no more likely to be published than were those with a negative outcome ($p = 0.410$). Studies with a negative outcome were of higher quality ($p = 0.003$) and included larger sample sizes ($p = 0.05$). Commercially funded ($p = 0.027$) and United States-based ($p = 0.020$) studies were more likely to be published, even though those studies were not associated with higher quality, larger sample sizes, or lower levels of evidence ($p = 0.24$ to 0.79).

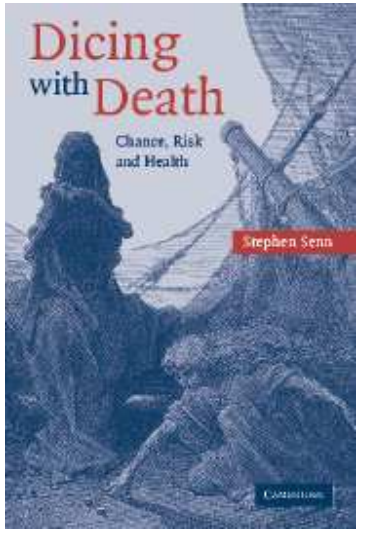
Conclusions

Relax. The lecture is nearly over

Statistics is more than just mathematics

- Basic philosophy is important
- Reaching for mathematics too quickly can harm understanding
  - Mathematical models are not the be all and end all
- It is important to see what the essence of a problem is
  - Graphics can help explain things to yourself and others
- Always ask
  - What will be known and what will be unknown?
  - How do I get to see what I see?
- And teach students this also!
- Understanding is necessary on more than one level
- Heuristics are valuable

That binary event



- Which is more likely:
  - You will witness exactly one million successes
  - You will witness 500,000 successes and 500,000 failures in any order?
- You can work with the predictive distribution of a beta-binomial if you like
  - Integration and algebra
- Or you can *see* the answer
- Given the prior after one million events the relative frequency will be the true value
- But your prior says every true value is equally likely
- Therefore the two events are equally likely

Final thought

Mathematics is full of lemmas but statistics is full of dilemmas

Some references

Senn, S. J. (1987). Correcting for regression in assessing the response to treatment in a selected population [letter]. *Statistics in Medicine, **6(6), 727-728***
Senn, S. J. (1998). Mathematics: governess or handmaiden? *Journal of the Royal Statistical Society Series D-The Statistician, **47(2), 251-259***
Senn, S. J. (1994). Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation [letter; comment] [see comments]. *Statistics in Medicine, **13(21), 2280-2285***
Senn, S. (2008). A note concerning a selection "Paradox" of Dawid's. *American Statistician, **62(3), 206-210***
Senn, S. (2014). A note regarding meta-analysis of sequential trials with stopping for efficacy. *Pharmaceutical statistics, **13(6), 371-375***
Senn, S. (2012). Misunderstanding publication bias: editors are not blameless after all. *F1000Research, **1***
Senn, S. (2013). Authors are also reviewers: problems in assigning cause for missing negative studies. *F1000Research. Retrieved from http://f1000research.com/articles/2-17/v1*

stephen.senn@lih.lu