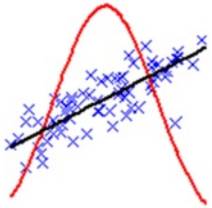
The replication crisis: are P-values the problem and are Bayes factors the solution?

Stephen Senn, Consultant Statistician, Edinburgh



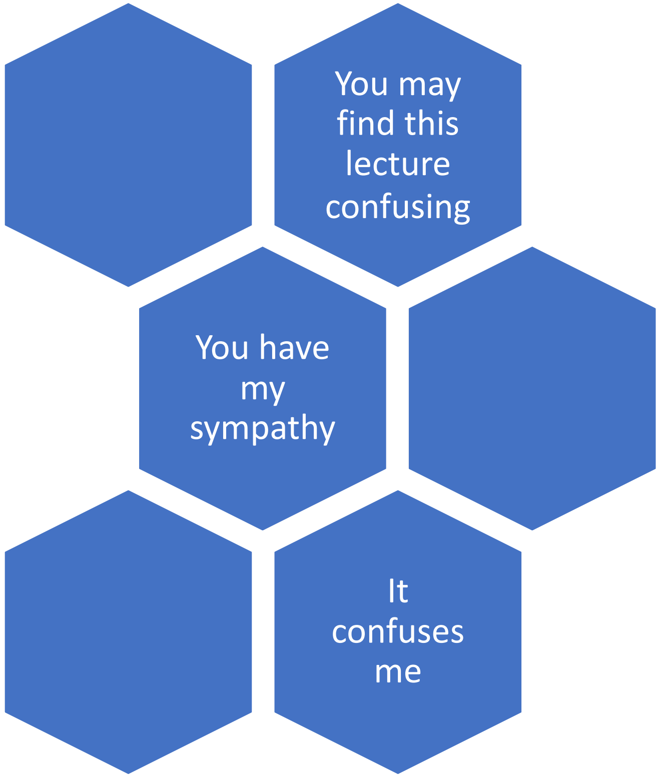
Phil-Stat Wars 2022

@stephensenn

stephen@senns.uk

Warning and apology

Warning



You may find this lecture confusing

You have my sympathy

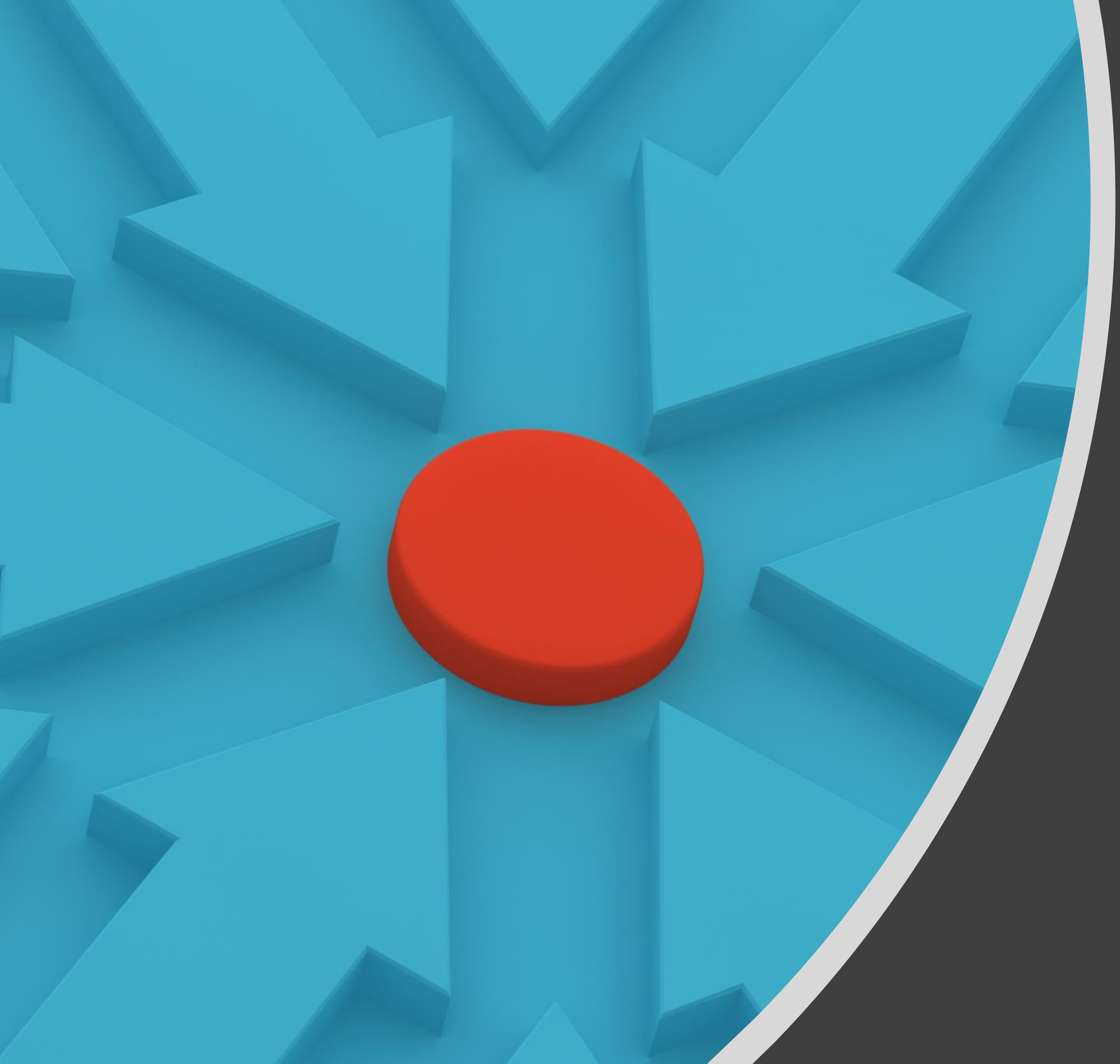
It confuses me

Apology

- This is an update of a lecture I gave in Zurich just before the pandemic (early 2020)
- I had expected to have had a lot more different things to say by now
- I was wrong

Outline

- What do we want from replication?
- P-values as inferential villains
- The confidence interval paradox
- Will Bayes factors proves superior
- Conclusions



What do we want from replication?

Repeat after me

Basic position

- A common current view is that we have a replication crisis
- To agree whether this is true or not depends (at least) on agreeing on what constitutes replication
- For example, are we concerned about:
  - Not replicating positive claims?
  - Not replicating negative claims?
- What do we want from the analysis of an experiment:
  - A statement of the evidence from that experiment?
  - A statement about the state of nature?
  - A prediction about a future experiment?
- In my opinion if there are problems, they have a lot less to do with choice of inferential system or statistic and a lot more to do with appropriate calculation

Theories of truth
Ralph C Walker

Some that are considered

- Correspondence
- Coherence
- Pragmatism
- Redundancy
- Semantic

https://onlinelibrary.wiley.com/doi/10.1002/9781118972090.ch21

Two I shall consider

- Correspondence
  - External consistency

  "The correspondence theory of truth holds that for a judgment… to be true is for it to correspond with the facts."
- Coherence
  - Internal consistency

  "The coherence theory of truth equates the truth of a judgment with its coherence with other beliefs."

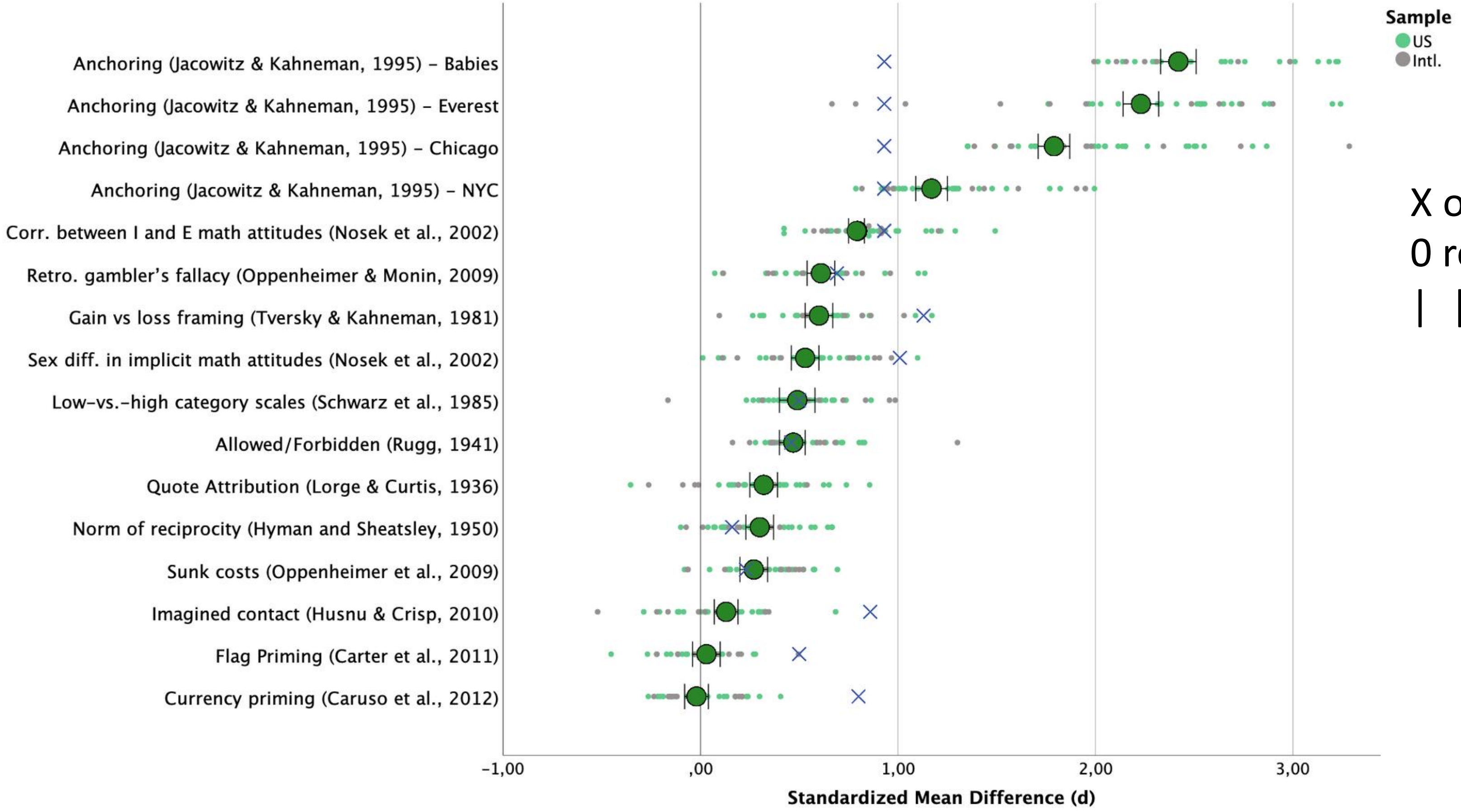
Many Labs Project

Thirty-six labs collaborate to check 13 earlier findings.
A large international group set up to test the reliability of psychology experiments has successfully reproduced the results of 10 out of 13 past experiments. The consortium also found that two effects could not be reproduced.

Ten of the effects were consistently replicated across different samples. These included classic results from economics Nobel laureate and psychologist Daniel Kahneman at Princeton University in New Jersey, such as gain-versus-loss framing, in which people are more prepared to take risks to avoid losses, rather than make gains1; and anchoring, an effect in which the first piece of information a person receives can introduce bias to later decisions. The team even showed that anchoring is substantially more powerful than Kahneman’s original study suggested.

Ed Yong
Yong, E. Psychologists strike a blow for reproducibility. Nature (2013).
https://doi.org/10.1038/nature.2013.14232

Investigating Variation in Replicability: A “Many Labs” Replication Project



X original result
0 replication
| | 99% error bars

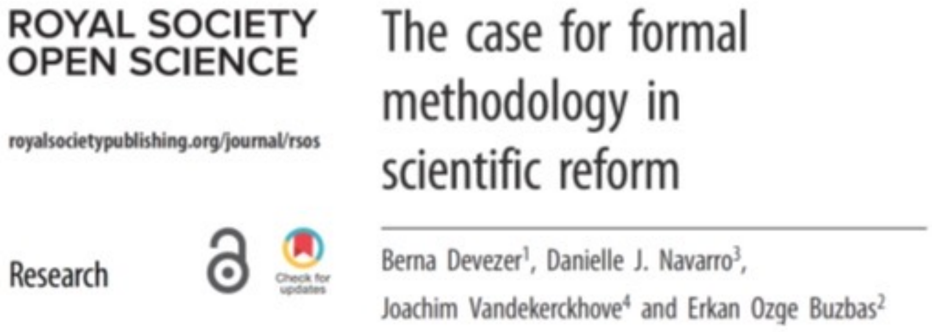
source
https://osf.io/WX7Ck/

Recommended reading

Points

- Reproducibility is a parameter of the population of studies
- **True results cannot be reproduced at will**
- **False results may be highly reproducible**
- Preregistration is not necessary for valid inferences

Paper



ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Check for updates

The case for formal methodology in scientific reform

Berna Devezer[1], Danielle J. Navarro[3], Joachim Vandekerckhove[4] and Erkan Ozge Buzbas[2]

Warning: some of these points need to be considered cautiously

P-values as inferential villains

Rejected for rejecting hypotheses

The villain in the story

- More and more statisticians (and others) are claiming that the P-value is the villain in the story
- And that RA Fisher is the evil mastermind behind the story
- This is odd, however, since
  - Fisher did not invent P-values
  - His major innovation (doubling them) was to make them more conservative
  - P-values are fairly similar to one standard Bayesian approach to inference*

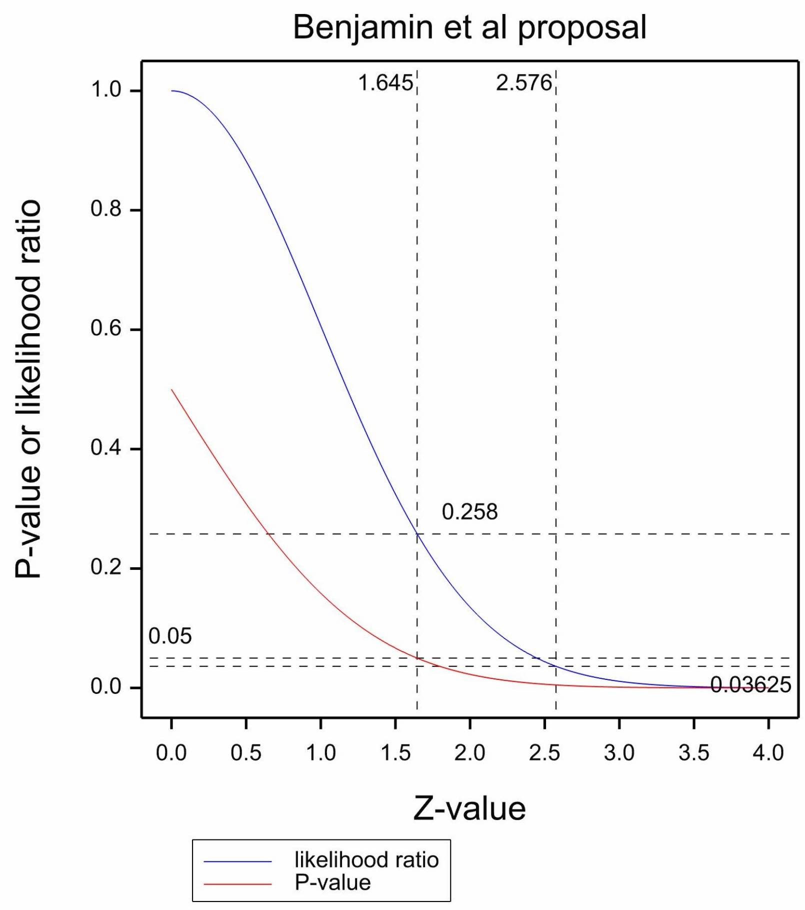
“For fields where the threshold for defining statistical significance for new discoveries is P < 0.05, we propose a change to P < 0.005. This simple step *would immediately improve the reproducibility of scientific research in many fields*.” (My emphasis.)

Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nature human behaviour*. 2018;2(1):6-10.

*The approach using ‘uninformative’ prior distributions and interpreting the one-sided P-value as the posterior probability that the apparently better treatment is actually worse.

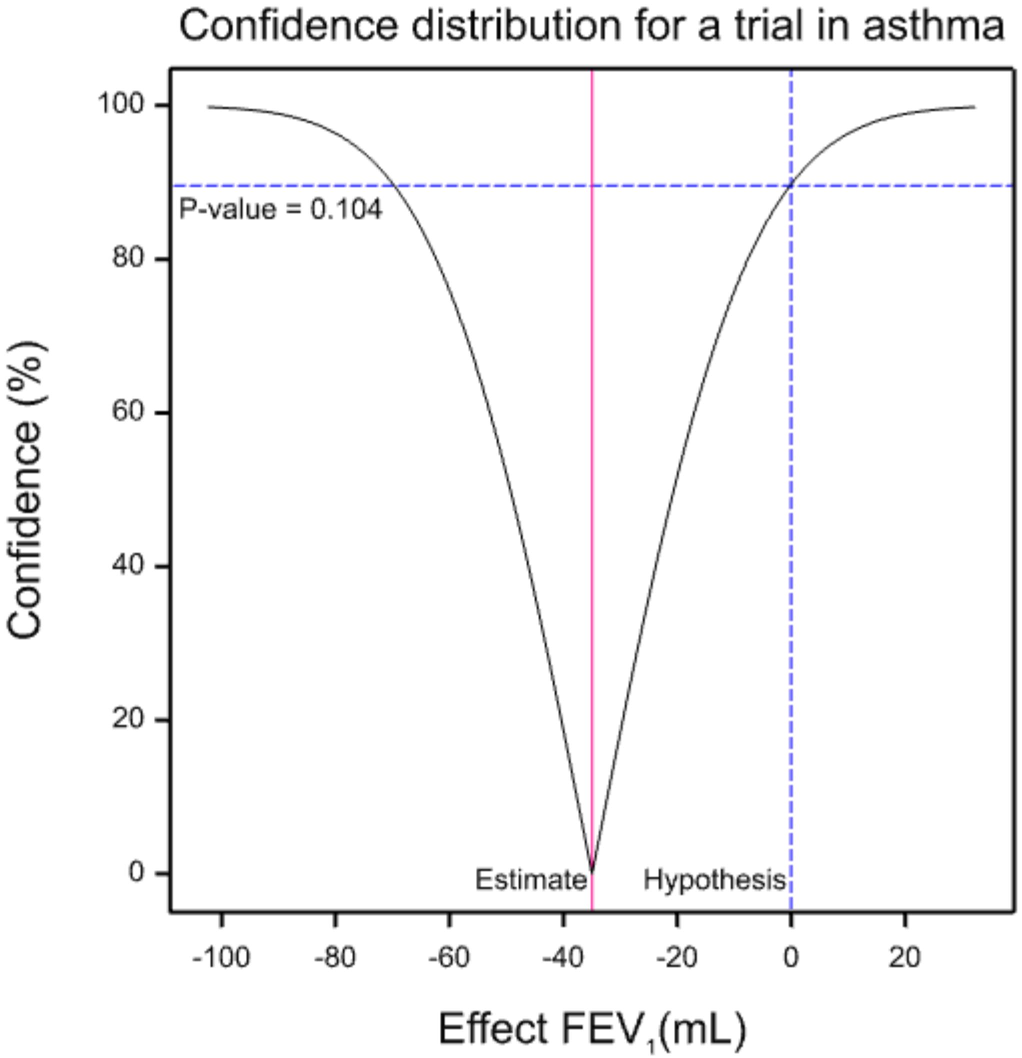
Significance level reform

- The argument is along the lines that a conventionally significant result (one sided) provides weak evidence in terms of likelihood ratio
- The stricter threshold provides much stronger evidence
  - LR is 0.036 rather than 0.258



From P-values to Confidence Intervals

- Everybody knows you can judge significance by looking at P-values
- Nearly everybody knows that you can judge significance by looking at confidence intervals
    - For example, does the confidence interval include the hypothesised value?
- Rather fewer people know that you can judge P-values by looking at confidence intervals
    - Just calculate lots of different degrees of confidence (a so-called confidence distribution) and see which one just excludes the hypothesised value
- It seems implausible, therefore, if P-values are inherently unreliable, confidence intervals must be reliable.
- So I am going to have a look at a confidence interval

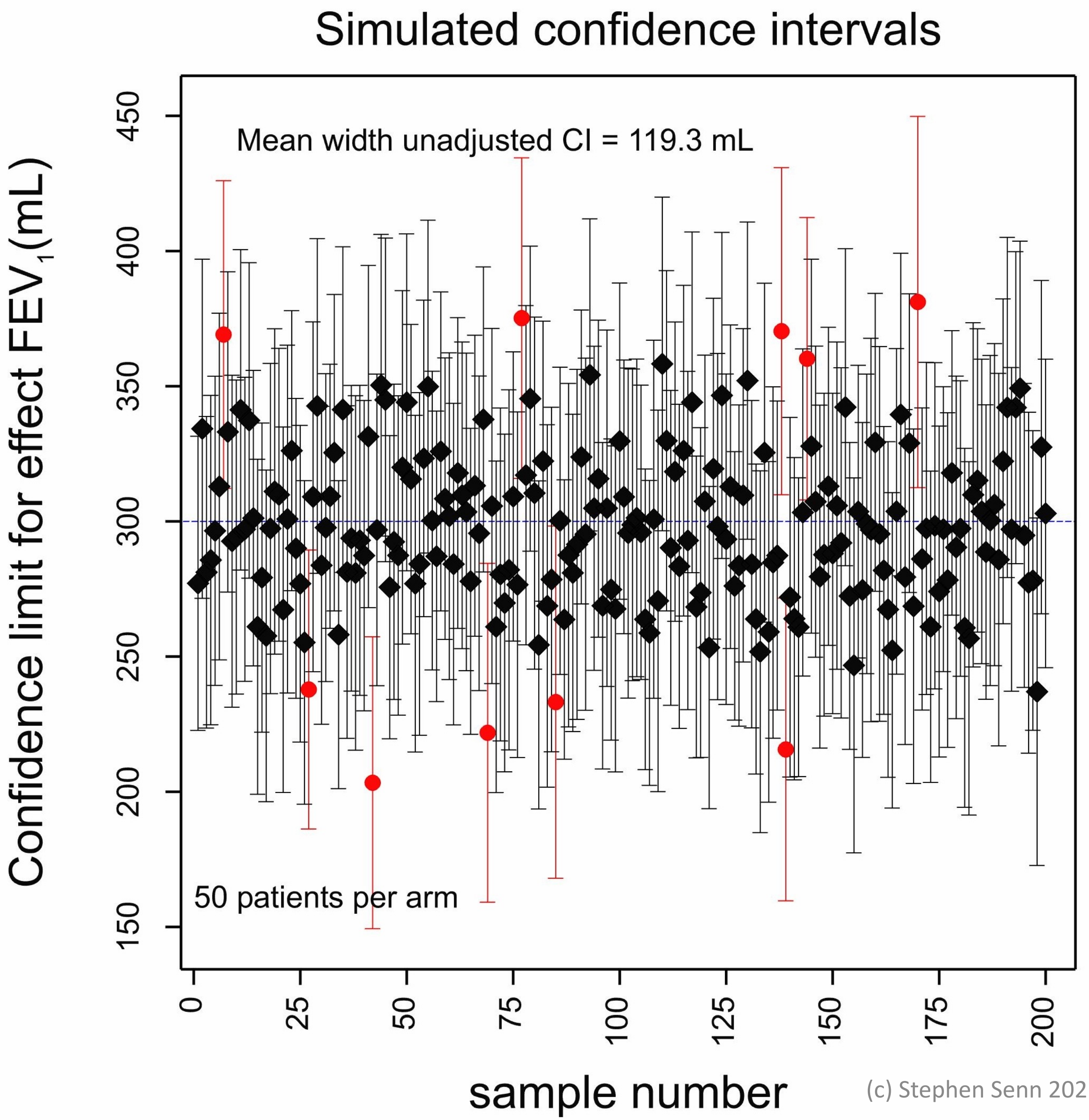


A trial in asthma with a disappointing result, which will be described as…

“Very nearly a trend towards significance”

The confidence interval paradox

Dual but cool?



Asthma: placebo controlled trials

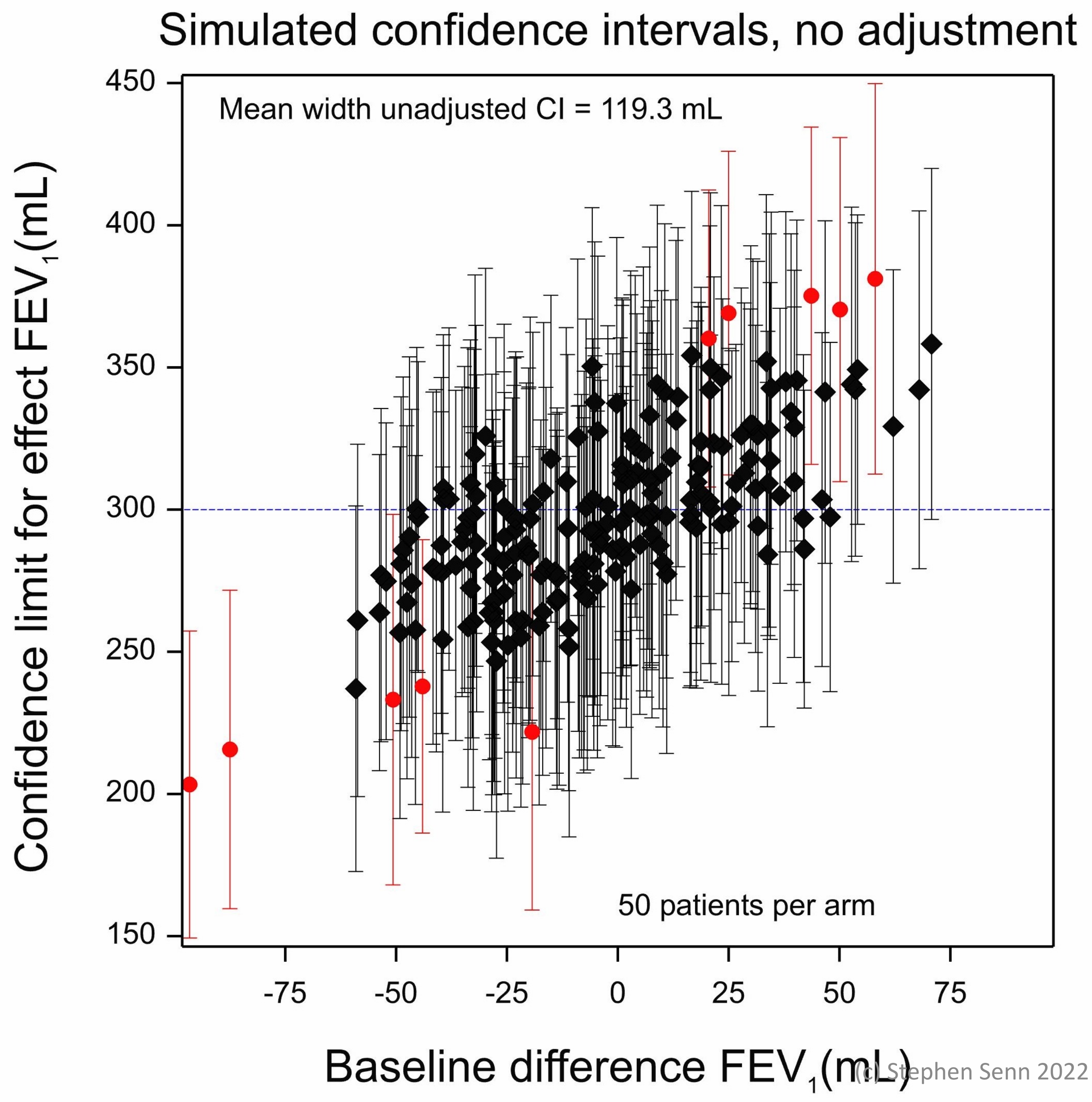
200 simulated confidence intervals where the true treatment effect is null

There are 200 intervals and 10 of them (in red) do *not* include the true value. (Therefore 190/200= 95% *do* include then true value.)

But what does this mean?

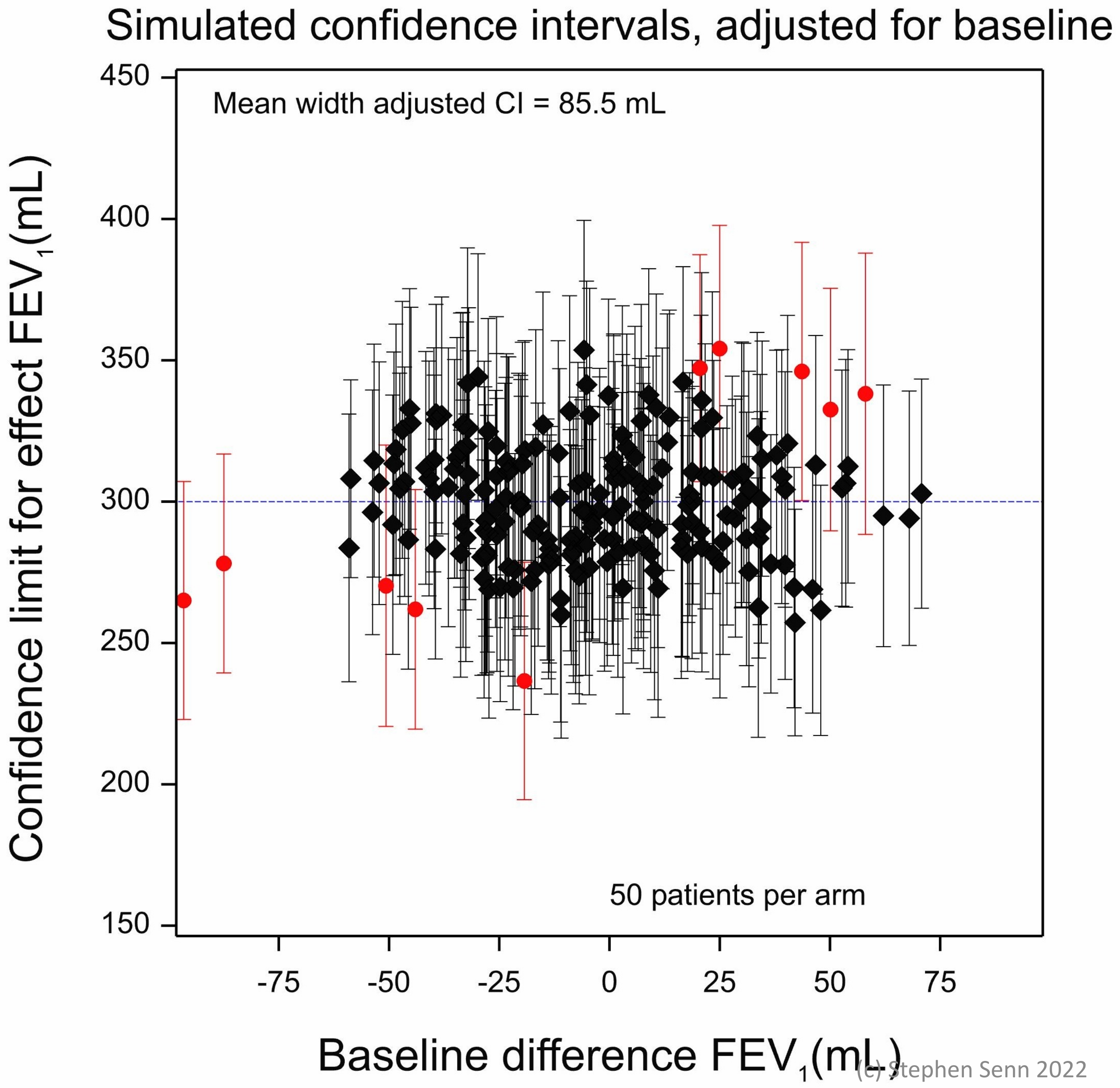
Let’s add some information.

(c) Stephen Senn 2022



Adding some information about baselines means that you can recognise intervals that are less likely to include the true value.

They are the ones at the extremes

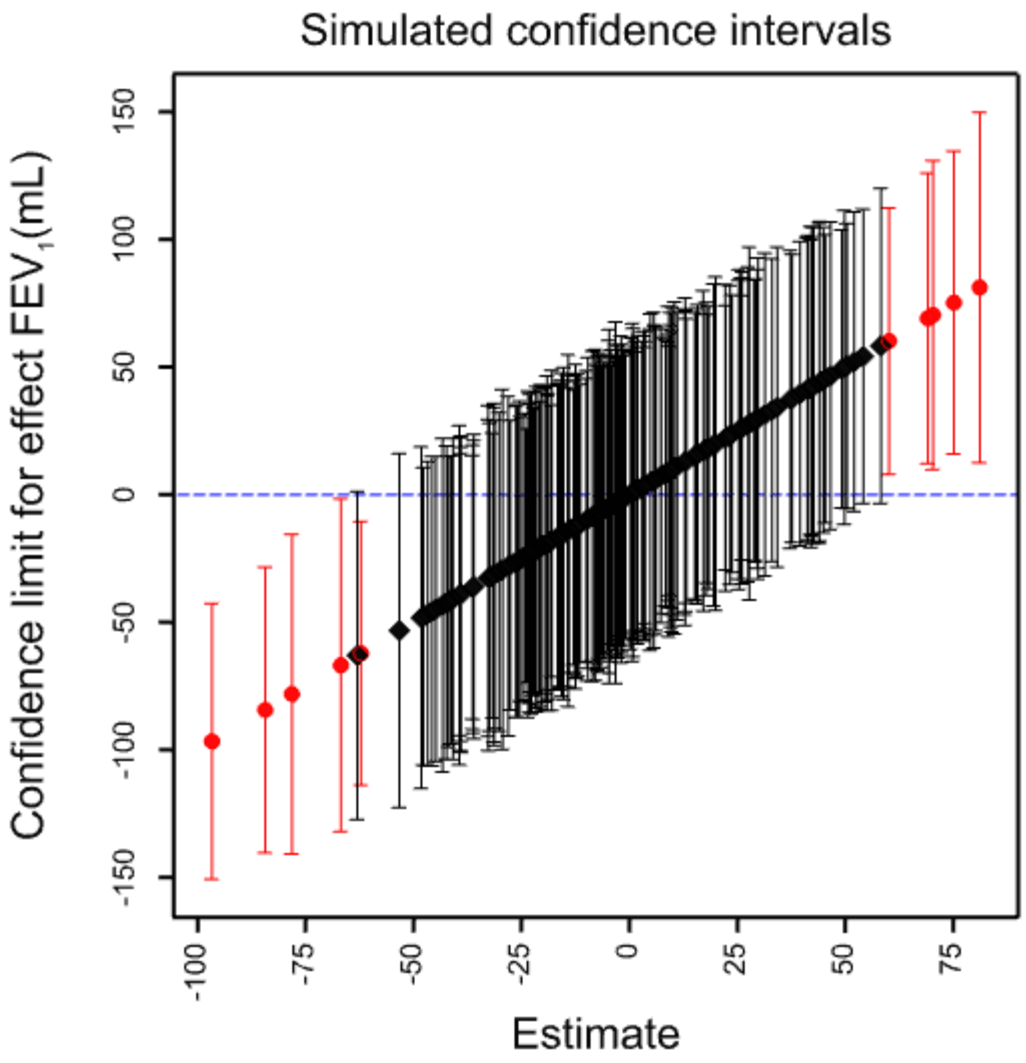


Conditioning on the baseline, however Cures the problem.

Again we have 10 intervals (5%) that do not include the true value

However, we can no longer see which intervals are unreliable

And there is an added bonus: we get narrower confidence intervals



But suppose we have no covariate information. How can we identify those unreliable intervals?

We could sort them by estimate as in the plot on the left.

There is only one snag. In practice we will only have one estimate and interval.

We can't compare point estimates to find those that are extreme.

Our only hope is to ***recognise*** whether or not they are extreme based on prior information.

This means being Bayesian.

So this explains the paradox

- The Bayesian can still maintain/agree these three things are true
  - P-values over-state the evidence against the null
  - P-values are 'dual' to confidence intervals
  - 95% of 95% confidence intervals will include the true value
- The reason is
  - Those 95% confidence intervals that *include* zero have a probability *greater* than 95% of including the true value
  - Those 95% intervals that *exclude* zero have a *less* than 95% chance of including the true value
  - On average it's OK
- The consequence is that your prior belief in zero as a special (much more probable value) gives you the means of recognising the less reliable intervals

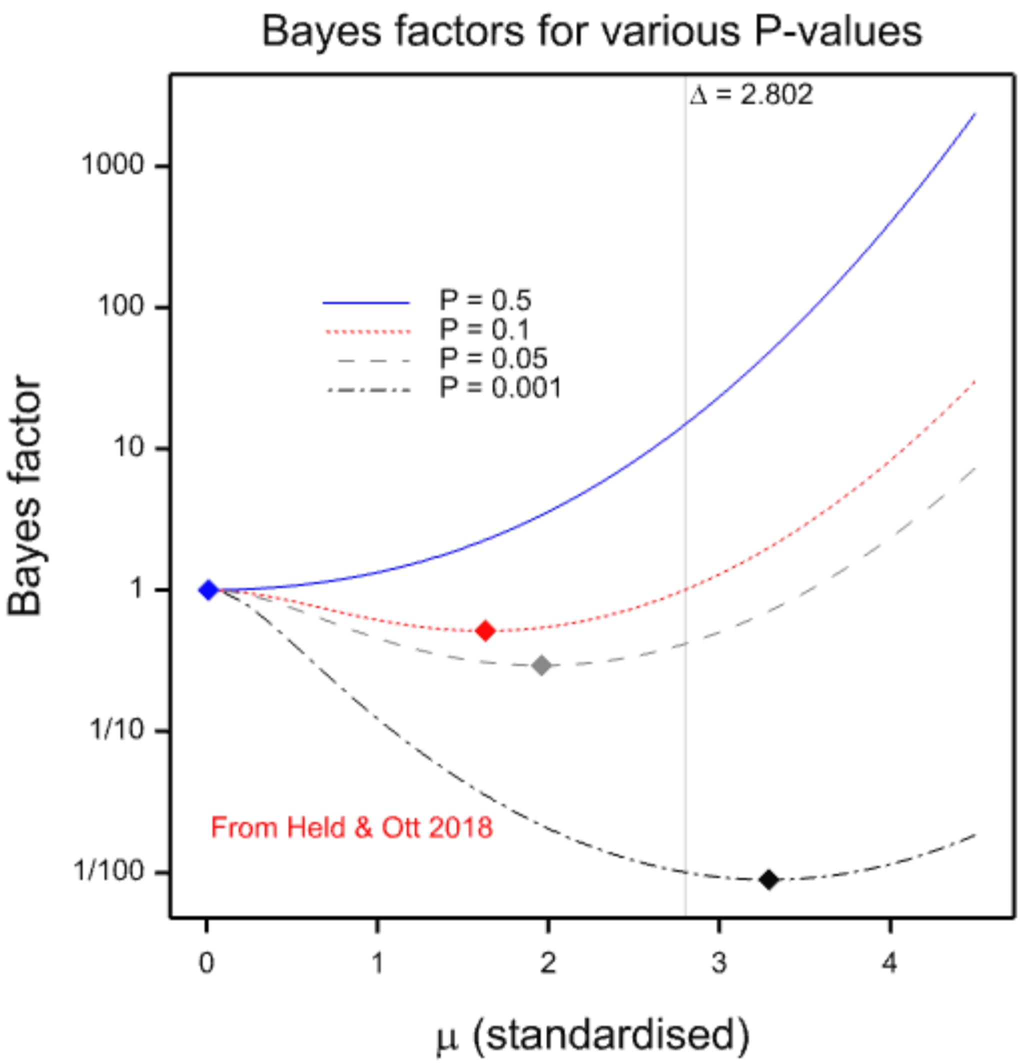
Will Bayes factors proves superior

Gained in translation?

Being Bayesian

This requires you to have a prior distribution

Approach	Advantages	Disadvantages
Formal	Automatic Software readily available JASP	No guarantees of anything Unlikely to reflect what you feel Unlikely to reflect what you know Unlikely to predict what will happen
Subjective	Never lose a bet with yourself Logically elegant	You can’t change your mind you can only update it
Knowledge based	May have some chance of working as an objectively testable prediction machine Useful for practical decision-making	Hard Very hard Extremely hard



This is merely one among many possibilities

The more traditional Jeffreys approach would generally produce more modest values

Note however, that if you have a dividing hypothesis and an uninformative prior, the one sided P-values can be given a Bayesian interpretation

Fisher simply gave a statistic that was commonly calculated and given a Bayesian interpretation, a different possible justification

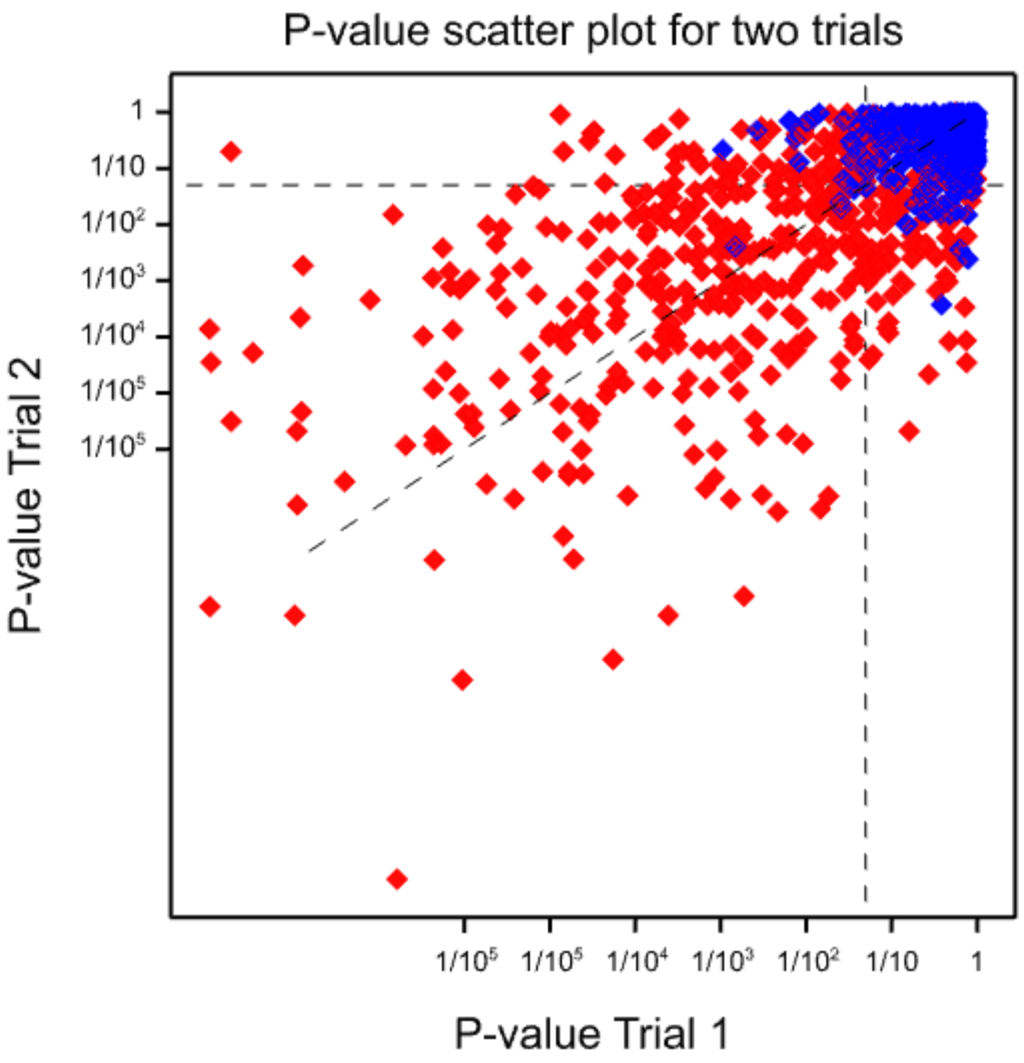
Goodman's Criticism

Sauce for the goose

- What is the probability of repeating a result that is just significant at the 5% level (p=0.05)?
- Answer 50%
    - If true difference is observed difference
    - If uninformative prior for true treatment effect
- Therefore, P-values are unreliable as inferential aids

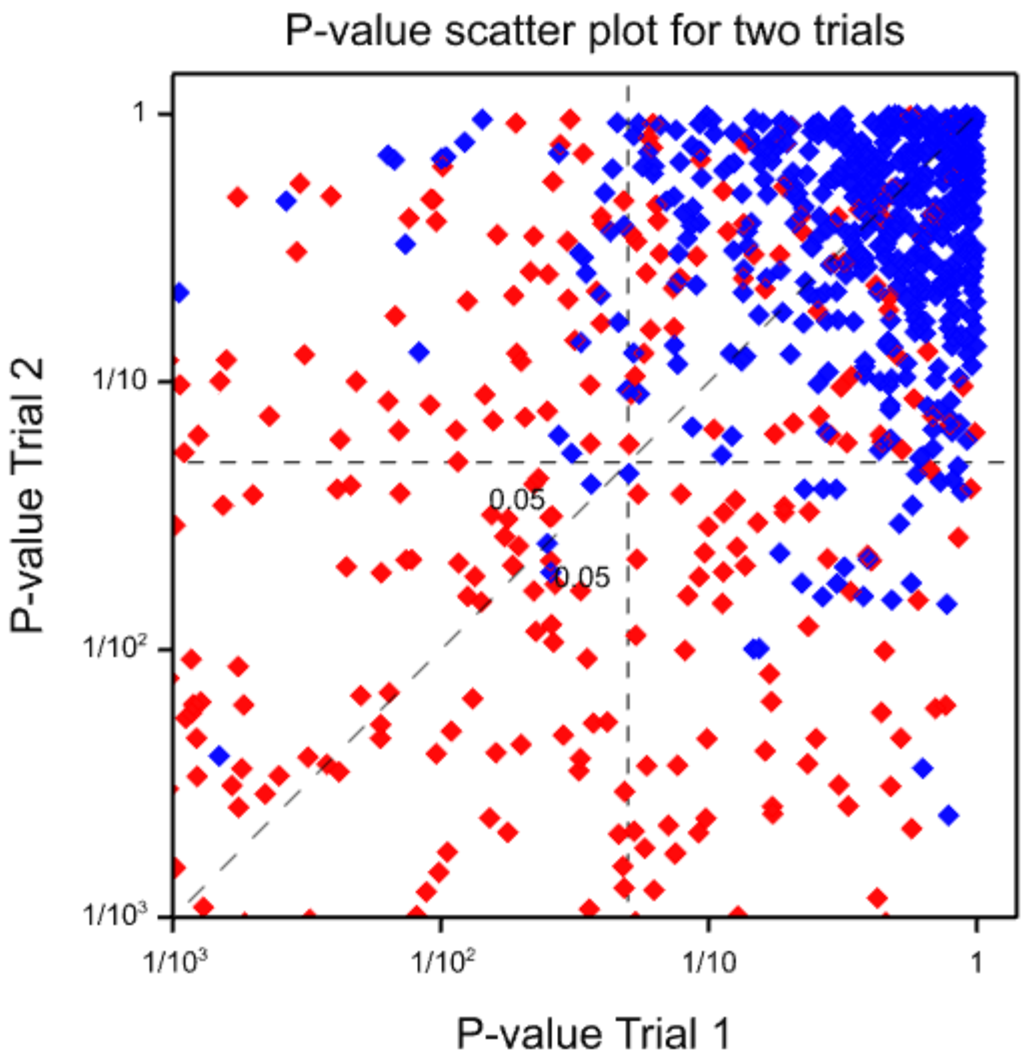
Sauce for the gander

- This property is shared by Bayesian statements
    - It follows from the Martingale property of Bayesian forecasts
- Hence, either
    - The property is undesirable and hence is a criticism of Bayesian methods also
    - Or it is desirable and is a point in favour of frequentist methods



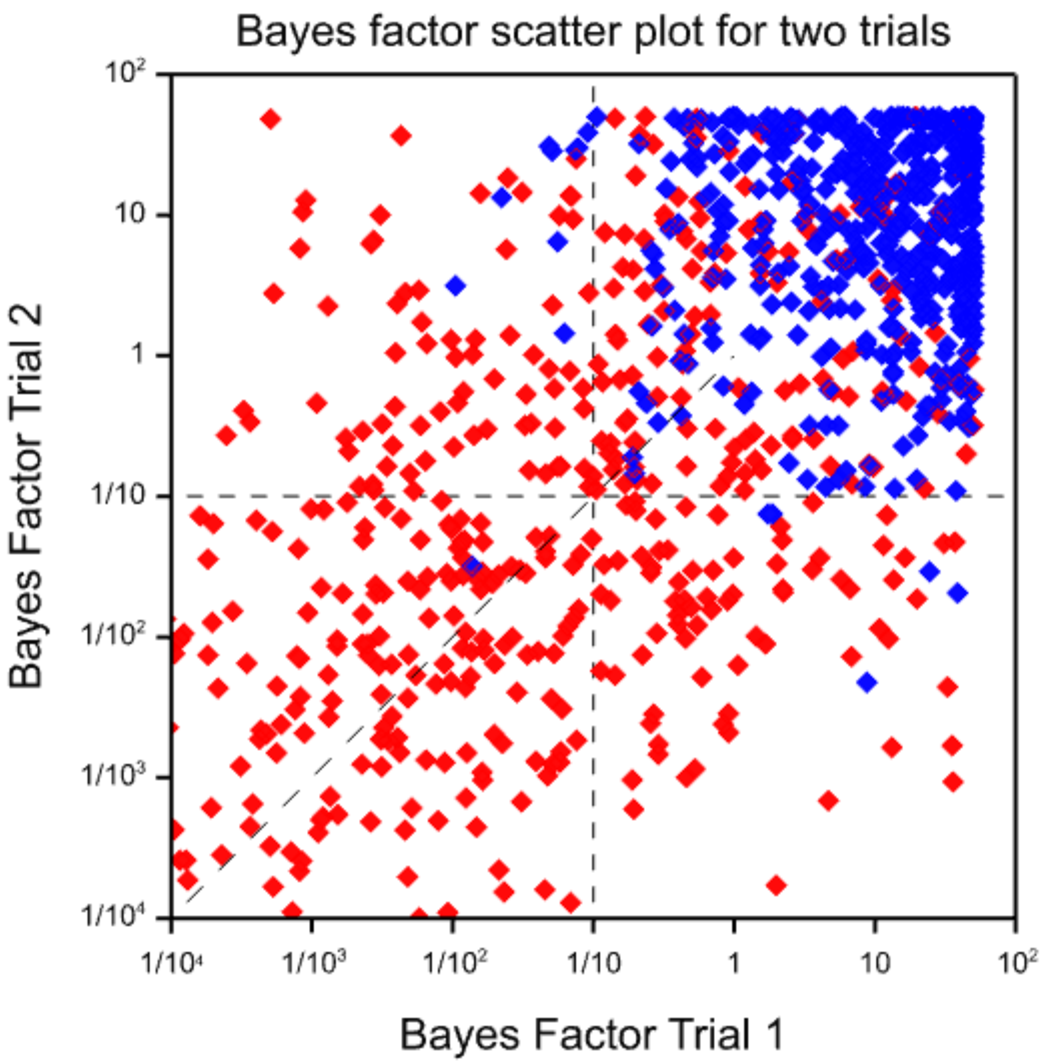
Parameter	H0	H1
Mean, μ	0	2.8
Variance of true effect	0	1
Standard error Of statistic	1	1

1000 trials. In half the cases H0 is true



As before but magnified to concentrate more easily on the 'significance' boundary of 0.05

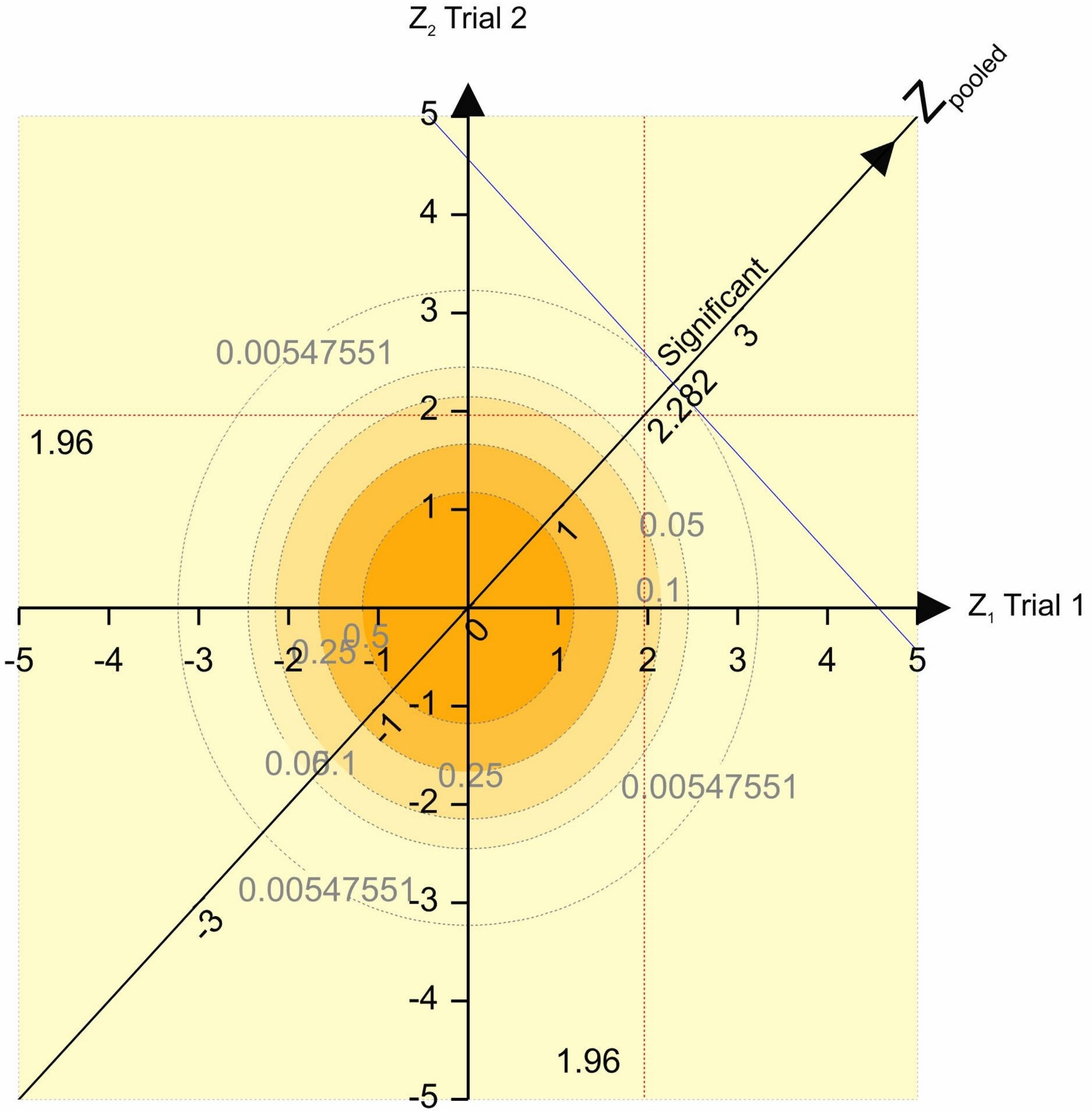
	Count		
2nd Significant	Yes	No	Count
1st Significant			
Yes	286	86	372
No	89	539	628
Count	375	625	1000



Magnified plot only

	Count		
2nd Strong	Yes	No	Count
1st Strong			
Yes	213	82	295
No	82	623	705
Count	295	705	1000

NB Bayes factor calculated at non-centrality of 2.8, which corresponds to the value for planning



A common regulatory standard is two trials significant at the 5% level

Since, in practice we are only interested in ‘positive’ significance, this is two trials significant one-sided at the 5% level

This diagram shows the ‘success space’ (to the right and above the red lines)

Replacing this by a meta-analysis (right and above the blue line) would require pooled significance at the $2\times\left(\frac{1}{40}\right)^2 = \frac{1}{800}$ level

The Z value corresponding to 1/1600 is 3.227 and dividing this by root 2 gives 2.282.

Comparison of Two Two-Trial Strategies

Simulation values are as previously

Second trial is only run if first is 'positive'

Efficacy declared if both trials are 'positive'

Figures are per 1000 trials for which H_0 is true in 500 and H_1 in 500

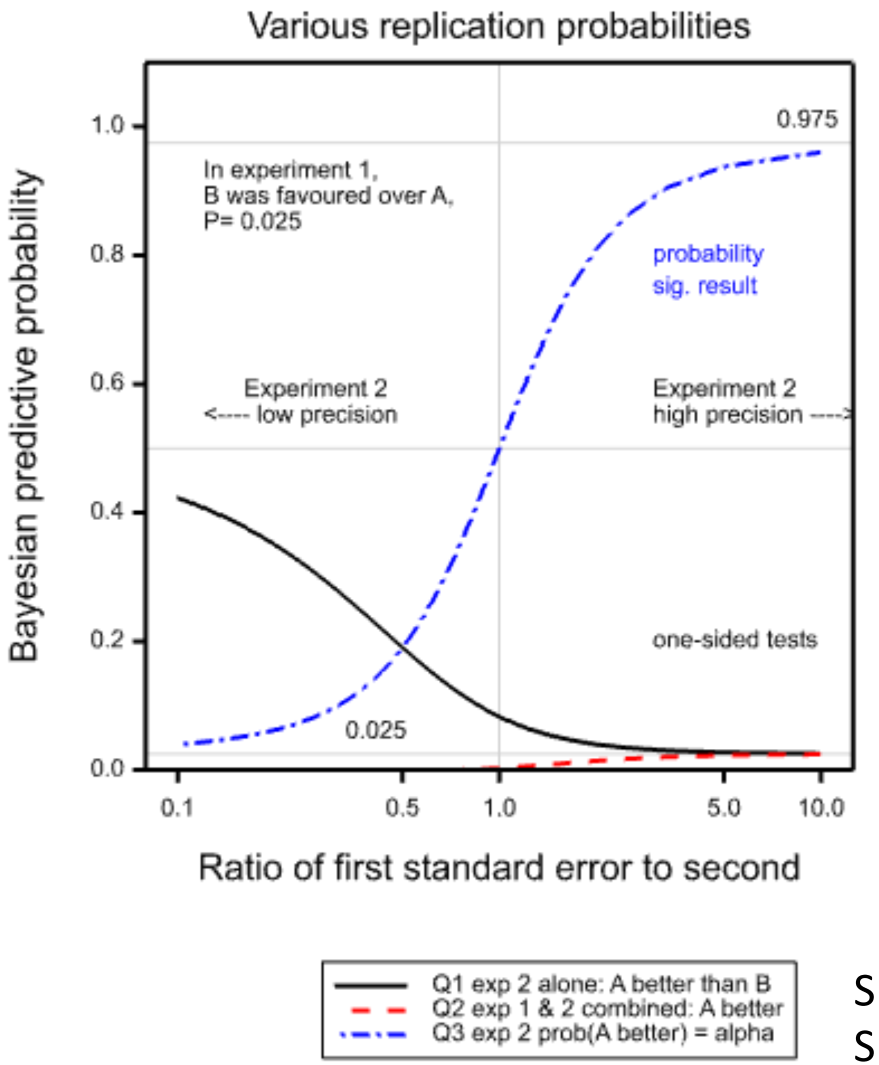
Statistics	Frequentist P<0.05	Bayesian BF<1/10
Number of trials run in total	1372	1299
Number of false positives	0	1
Number of false negatives	68	133
Number of false decisions	68	134
Number of correct decisions	932	866
Trials per correct decision	1.47	1.5

So what does this amount to?

- Nothing more than this
    - Given the same statistic calculated for two exchangeable studies then, in advance, the probability that the first is higher than the second is the same as the probability that the second is higher than the first
    - Once a study has reported, for any bet to be more than a coin toss, other information has to be available
        - For example that, given prior knowledge, the first is implausibly low
- In *theory*, a proper Bayesian analysis exhausts such prior knowledge so no further adjustment of what is known on average is necessary
- One has to be very careful in using Bayes Factors
- One has to be very careful in specifying prior distributions

In a first experiment B was better than A (P=0.025) Three Possible Questions

- Q1 What is the probability that in a future experiment, taking that experiment's results *alone*, the *estimate* for B would after all be worse than that for A?
- Q2 What is the probability, having conducted this experiment, and *pooled* its results with the current one, we would show that the *estimate* for B was, after all, worse than that for A?
- Q3 What is the probability that having conducted a future experiment and then calculated a Bayesian posterior using a uniform prior and the results of this second experiment *alone*, the *probability* that B would be worse than A would be less than or equal to 0.05?



Q3 is the repetition of the P-value.
Its probability is ½ provided that the experiment is exactly the same size as the first.

If the second experiment is larger than the first, this probability is greater than 0.5

If it is smaller that the first, this probability is smaller than 0.5.

In the limit as the size goes to infinity, the probability reaches 0.975.

That’s because the second study now delivers the truth and that, if we are going to be Bayesian with an uninformative prior is what the posterior is supposed to deliver

S. J. Senn (2002) A comment on replication, p-values and evidence S.N.Goodman, Statistics in Medicine 1992; 11:875-879. Statistics in Medicine,21, 2437-2444.

What are we really interested in?

- We are interested in the truth
- We are not interested in ‘replication’ *per se*
  - Certainly not if the replication study is no larger than the original study
  - This replication probability is irrelevant
- We are interested in what an infinitely large study would show
- This suggests that what we are really interested in is meta-analysis as a way of understanding results
- This is actually rather Bayesian

Benjamin et al revisited

- So how can it be possible that using a new standard for significance would improve replication
- Answer: it can’t
- Benjamin et al cheat
- They keep the *old* standard for replication
  - P<0.005 is to be <u>replicated</u> by P<0.05
- But it is far from clear that in a quest for true results, rather than merely replicated results, the strategy of requiring impressive initial results and moderate confirmation, is better than vice versa

“Empirical evidence from recent replication projects in psychology and experimental economics provide insights into the prior odds in favour of H1. In both projects, the rate of replication (that is, significance at P < 0.05 in the replication in a consistent direction) was roughly double for initial studies with P < 0.005 relative to initial studies with 0.005 < P < 0.05: 50% versus 24% for psychology”, and 85% versus 44% for experimental economics”

Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nature human behaviour*. 2018;2(1):6-10.

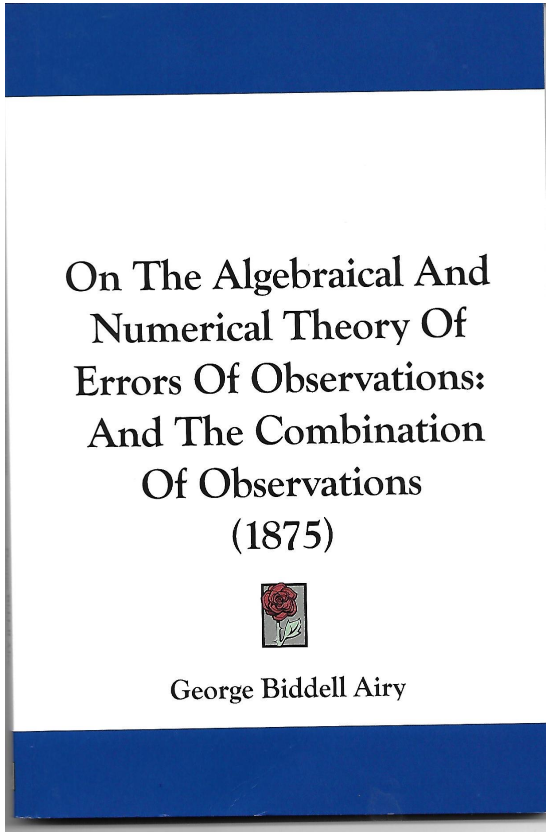
Lessons

- This does <u>not</u> show that significance is better than Bayes factors
  - Different parameter settings could show the opposite
  - Different thresholds could show the opposite
- It does show that reproducibility is not necessarily the issue
  - The strategy with poorer reproducibility is (marginally) better
    - NB That does not have to be the case
    - But it shows it can be the case
- It raises the issue of calibration
- And it also raises the issue as to whether reproducibility is a useful goal in itself
  - And how it should be measured

George Biddell Airy

1801-1892

Astronomer Royal from 1835

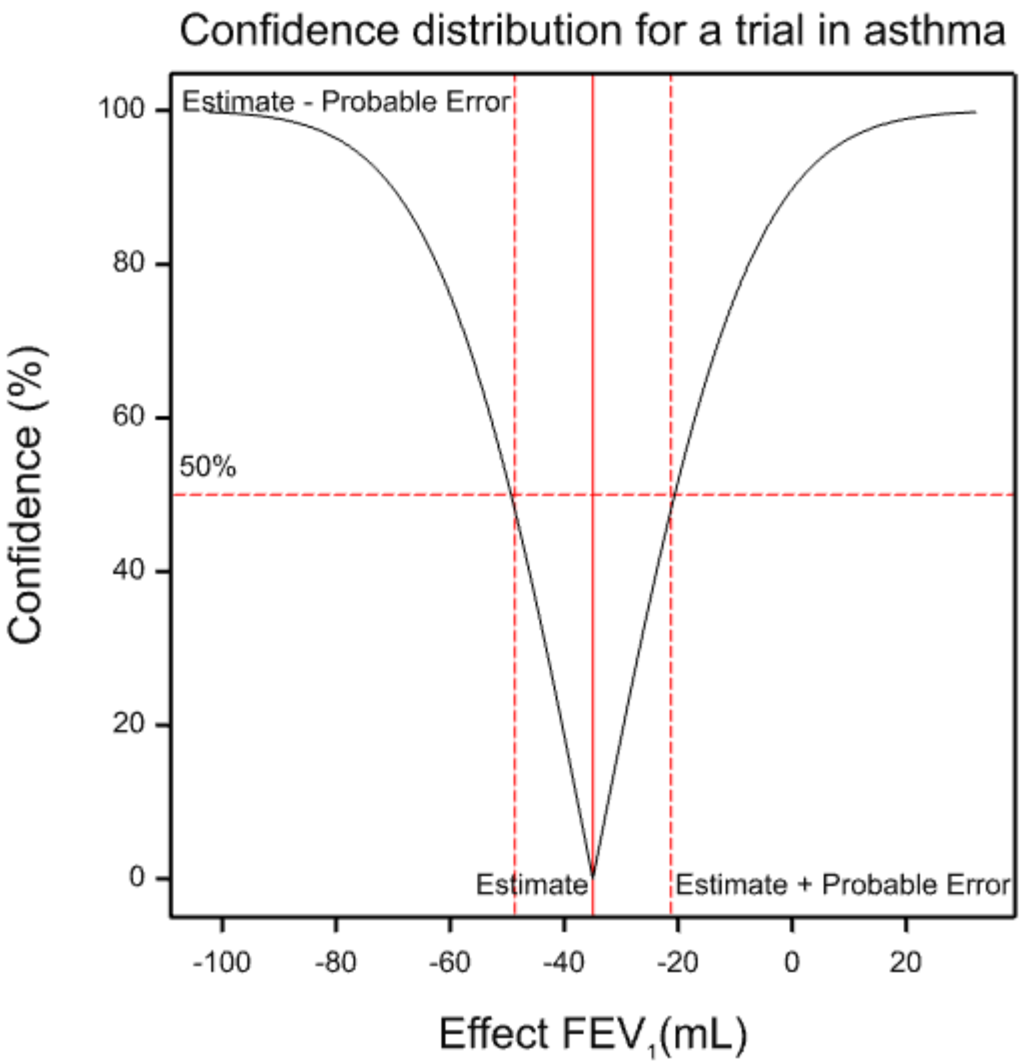


On The Algebraical And Numerical Theory Of Errors Of Observations: And The Combination Of Observations (1875)

George Biddell Airy

First edition 1866

This was a book that Student mentioned in correspondence with Fisher



$$t.w. = 1/(probable\ error)^2$$

Probable error ≅ 0.645 x SE

First. The combination-weight for each measure ought to be proportional to its theoretical weight.

Second. When the combination-weight for each measure is proportional to its theoretical weight, the theoretical weight of the final result is equal to the sum of the theoretical weights of the several collateral measures. (pp55-56).

$$\text{Estimate}_{posterior} = \frac{\text{Precision}_{prior} \times \text{Estimate}_{prior} + \text{Precision}_{data} \times \text{Estimate}_{data}}{\text{Precision}_{prior} + \text{Precision}_{data}}$$

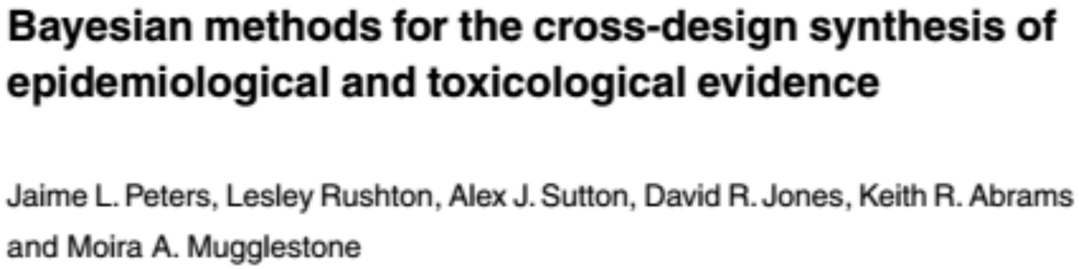
$$\text{Precision}_{posterior} = \text{Precision}_{prior} + \text{Precision}_{data}$$

Summing up

- The connection between P-values and Bayesian inference is very old
- The idea of carrying out frequentist style fixed effects meta-analysis (but giving it a Bayesian interpretation) is very old
- The more modern idea that P-values somehow give significance too easily is related to using a different sort of prior distribution altogether
- It may be useful on occasion but it requires justification
- And in any case it ignores the problem of false negatives
- I am unenthusiastic about automatic Bayes
- We need many ways of looking at data
- But there are problems of data analysis we need to take seriously that transcend choice of inferential framework ….

An example

- Peters et al 2005, describe ways of combining different epidemiological (humans) and toxicological studies (animals) using Bayesian approaches
- Trihalomethanes and low birthweight used as an illustration
- However, their analysis of the toxicology data from rats treats the pups as independent, which makes very strong (implicit) and implausible assumptions
    - Pseudoreplication



Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence

Jaime L. Peters, Lesley Rushton, Alex J. Sutton, David R. Jones, Keith R. Abrams and Moira A. Mugglestone

Discussed in Grieve and Senn (2005) (Unpublished)

Standard errors reported as being smaller than they are

Block main effect problems

- Incorrect unit of experimental inference (Pseudoreplication, Hurlbert, 1984)
  - For example allocation of treatments to cages but counting rats as the ‘n’
- Inappropriate experimental analogue for observational studies
  - Historical data treated as if they were concurrent
    - Means parallel group trial is used as inappropriate model rather than cluster randomised trial
  - This may apply more widely to epidemiological studies

Block interaction problems

- Mistaking causal inference for predictive inference
  - Causal: What happened to the patients in this trial?
    - Interaction does not contribute to the error
      - Fixed effects meta-analysis is an example
  - Predictive what will happen to patients in future?
    - Interaction should contribute to the error
      - Random effects meta-analysis is an example

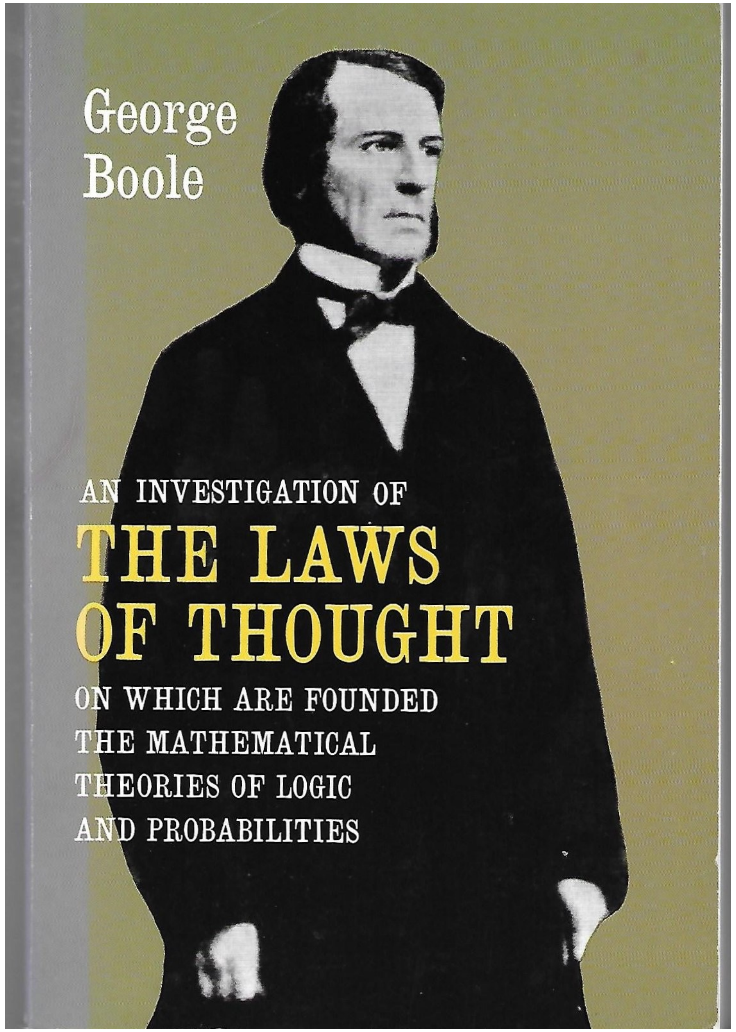
Other Problems

- Publication bias
- Hacking
- Multiplicity
  - In the sense of only concentrating on the good news
- Surrogate endpoints
- Naïve expectations

My ‘Conclusions’

- P-values *per se* are not the problem
- There may be a harmful culture of ‘significance’ however this is defined
- P-values have a limited use as rough and ready tools using little structure
- Where you have more structure you can often do better
  - Likelihood, Bayes etc
  - Point estimates and standard errors are extremely useful for future research synthesizers and should be provided regularly
- We need to make sure that estimates are reported honestly and standard errors calculate appropriately
- We should not confuse the evidence from a study with the conclusion that should be made

Finally, a thought about laws (or hypotheses) from *The Laws of Thought*



When the defect of data is supplied by hypothesis, the solutions will, in general, vary with the nature of hypotheses assumed.

Boole (1854) *An Investigation of the Laws of Thought*: Macmillan.
. P375
Quoted by R. A. Fisher (1956) Statistical methods and scientific inference In Statistical Methods, Experimental Design and Scientific Inference (ed J. H. Bennet), Oxford: Oxford University.

In worrying about the hypotheses, let's not overlook the defects of data

Where to find some of this stuff

http://www.senns.demon.co.uk/Blogs.html