P Values: the problem is not what you think

Stephen Senn

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Acknowledgements

Acknowledgements

Thanks to you all for inviting me and to Jarl Kampen for organizing it

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"

Outline

- A simple problem
  - The Bayesian approach & the frequentist P-value approach
- What is a P-value?
  - Beware of invalid inversion
- The crisis of replication
  - P-values as public enemy number one
- A brief history of P-values
  - The skeleton in the Bayesian cupboard
- What are we looking for in replication?
  - Repeat after me
- Empirical evidence
  - A dram of data is worth a pint of pontification
- Conclusions

An Example

My compact disc (CD) player* allowed me to press tracks in sequential order by pressing ***play*** or in random order by playing ***shuffle***.

One day I was playing the CD *Hysteria* by Def Leppard. This CD has 12 tracks.

I thought that I had pressed the *shuffle* button but the first track played was ‘women’, which is the first track on the CD.

Q. What is the probability that I did, in fact, press the *shuffle* button as intended?

*I now have an Ipod nano

An Apology

- You may find this example trivial, irrelevant and uninteresting but it has the following advantages.
  - It allows me to illustrate the major differences between two schools of statistics.
  - It is genuine.
  - simpler than standard genuine examples occurring in clinical research
  - I like it

Back to the Heavy Metal (The Bayesian Solution)

We have two basic hypotheses:

1) I pressed *shuffle*.

2) I pressed *play*.

First we have to establish a so-called ***prior probability*** for these hypotheses: a probability before seeing the evidence.

Suppose that the probability that I press the *shuffle* button when I mean to press the shuffle button is 9/10. The probability of making a mistake and pressing the *play* button is then 1/10.

Next we establish probabilities of events *given* theories. These particular sorts of probabilities are referred to as ***likelihoods*** ,a term due to RA Fisher(1890-1962).

If I pressed *shuffle*, then the probability that the first track will be ‘women’ (W) is 1/12. If I pressed *play*, then the probability that the first track is W is 1.

For completeness (although it is not necessary for the solution) we consider the likelihoods had any other track apart from ‘women’ (say X) been played.

If I pressed *shuffle* then the probability of X is 11/12. If I pressed *play* then this probability is 0.

We can put this together as follows

Hypothesis	Prior Probability P	Evidence	Likelihood	P x L
Shuffle	9/10	W	1/12	9/120
Shuffle	9/10	X	11/12	99/120
Play	1/10	W	1	12/120
Play	1/10	X	0	0
TOTAL				120/120 = 1

After seeing (hearing) the evidence, however, only two rows remain

Hypothesis	Prior Probability P	Evidence	Likelihood	P x L
Shuffle	9/10	W	1/12	9/120
Play	1/10	W	1	12/120
TOTAL				21/120

The probabilities of the two cases which remain do not add up to 1.

However, since these two cases cover all the possibilities which remain, their combined probability *must* be 1.

Therefore we rescale the individual probabilities to make them add to 1.

We can do this without changing their relative value by dividing by their total, 21/120.

This has been done in the table below.

So we rescale by dividing by the total probability

Hypothesis	Prior Probability P	Evidence	Likelihood	P x L	Posterior Probability
Shuffle	9/10	W	1/12	9/120	(9/120)/(21/120) =9/21
Play	1/10	W	1	12/120	(12/120)/(21/120) =12/21
TOTAL				21/120	21/21=1

The* posterior probability that I pressed shuffle is 9/21.

This completes the Bayesian solution.

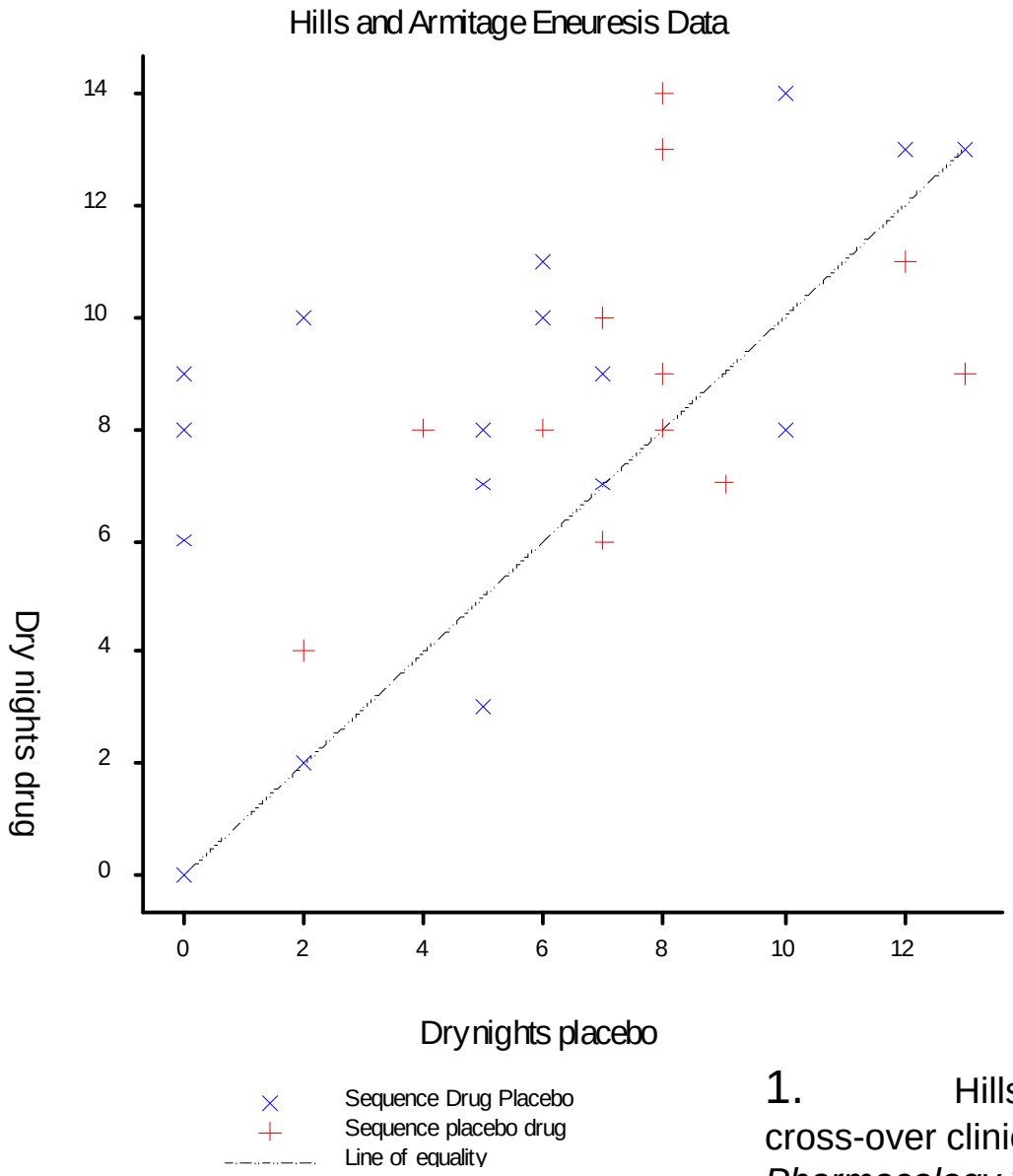
*Strictly speaking, **my** posterior probability.

Why don't we always do this?

- The calculation had to start with a subjective prior probability
- Some people don't like this but if you want to be Bayesian you can't avoid it
- The alternative is only to quote the evidence of the data given the hypothesis
- If I pressed shuffle the probability of Women being the first track played is 1/12
- This is the P-value

How to Calculate a P-value
An Example

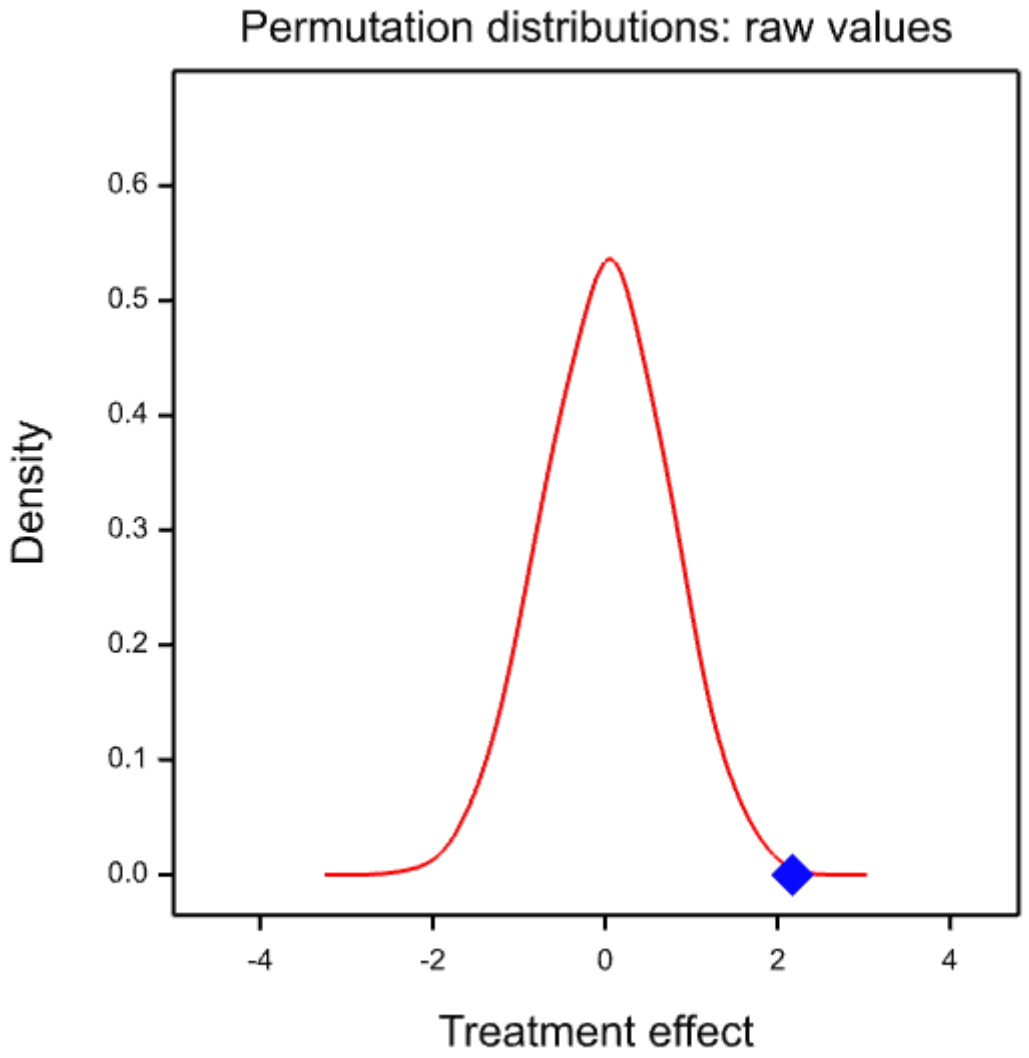
- Cross-over trial reported by Hills and Armitage
- Trial of a treatment in enuresis (bed-wetting)
- Patients randomised to one of two sequences
  - Treatment – placebo
  - Placebo-treatment
- Two weeks under each treatment
- Outcome variable is number of dry nights



Cross-over trial in Eneuresis

Two treatment periods of 14 days each

1. Hills, M, Armitage, P. The two-period cross-over clinical trial, *British Journal of Clinical Pharmacology* 1979; **8**: 7-20.


Permutation distributions: raw values
Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
Treatment effect

Beware of Invalid Inversion

- Is the Pope a Catholic?
  - Yes
- Is a Catholic the Pope?
  - Almost certainly not
- Assuming that the probability of A given B is the same as B given A is *invalid inversion*
- A P-value is a statement about the probability of (a certain aspect of) the data given the hypothesis
- It is not the probability of the hypothesis given the data

The Crisis of Replication

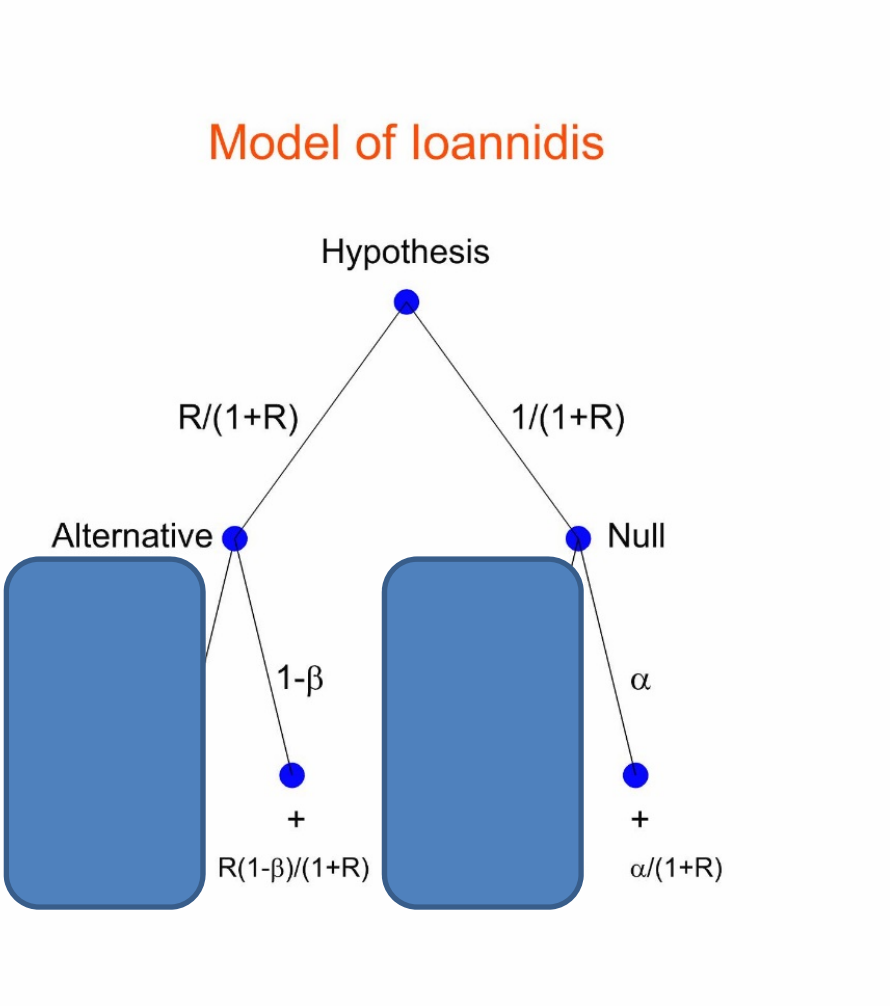
In countless tweets….The “replication police” were described as “shameless little bullies,” “self-righteous, self-appointed sheriffs” engaged in a process “clearly not designed to find truth,” “second stringers” who were incapable of making novel contributions of their own to the literature, and—most succinctly—“assholes.”

Why Psychologists’ Food Fight Matters
“Important findings” haven’t been replicated, and science may have to change its ways.

By Michelle N. Meyer and Christopher Chabris , *Science*

Ioannidis (2005)

- Claimed that most published research findings are wrong
  - By *finding* he means a 'positive' result
- 3569 citations by 21 April 2016 according to Google Scholar



$$TPR = \frac{R(1-\beta)/(1+R)}{[\alpha+R(1-\beta)]/(1+R)} = \frac{R(1-\beta)}{\alpha+R(1-\beta)}$$

(TPR=*True Positive Rate*)

Colquhoun's Criticisms

One must admit, however reluctantly, that despite the huge contributions that Ronald Fisher made to statistics, there is an element of truth in the conclusion of a perspicacious journalist:

> The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug. Robert Matthews [21] *Sunday Telegraph*, 13 September 1998.

"If you want to avoid making a fool of yourself very often, do not regard anything greater than $p < 0.001$ as a demonstration that you have discovered something. Or, slightly less stringently, use a three-sigma rule."

Royal Society Open Science 2014

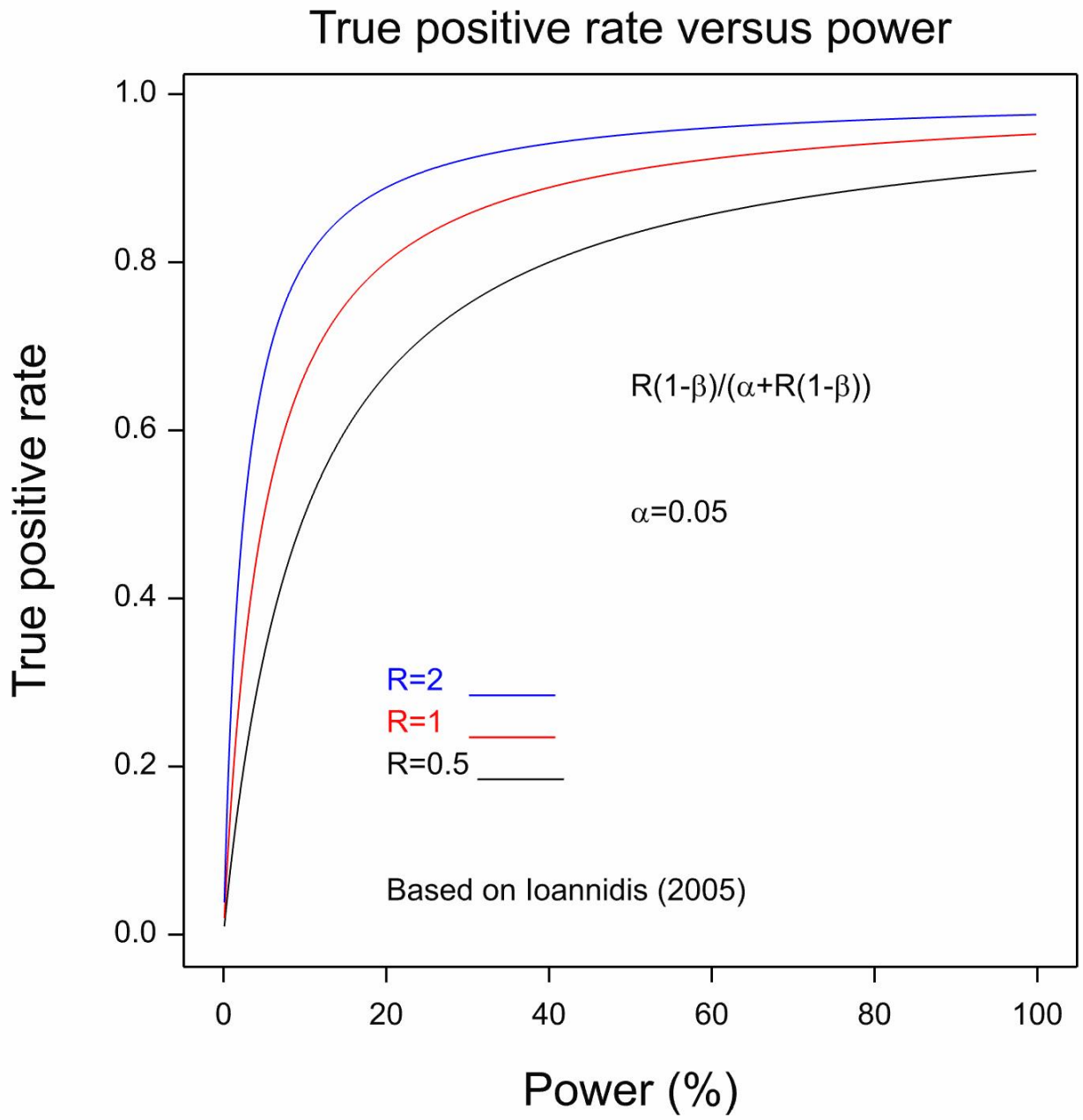
On the other hand

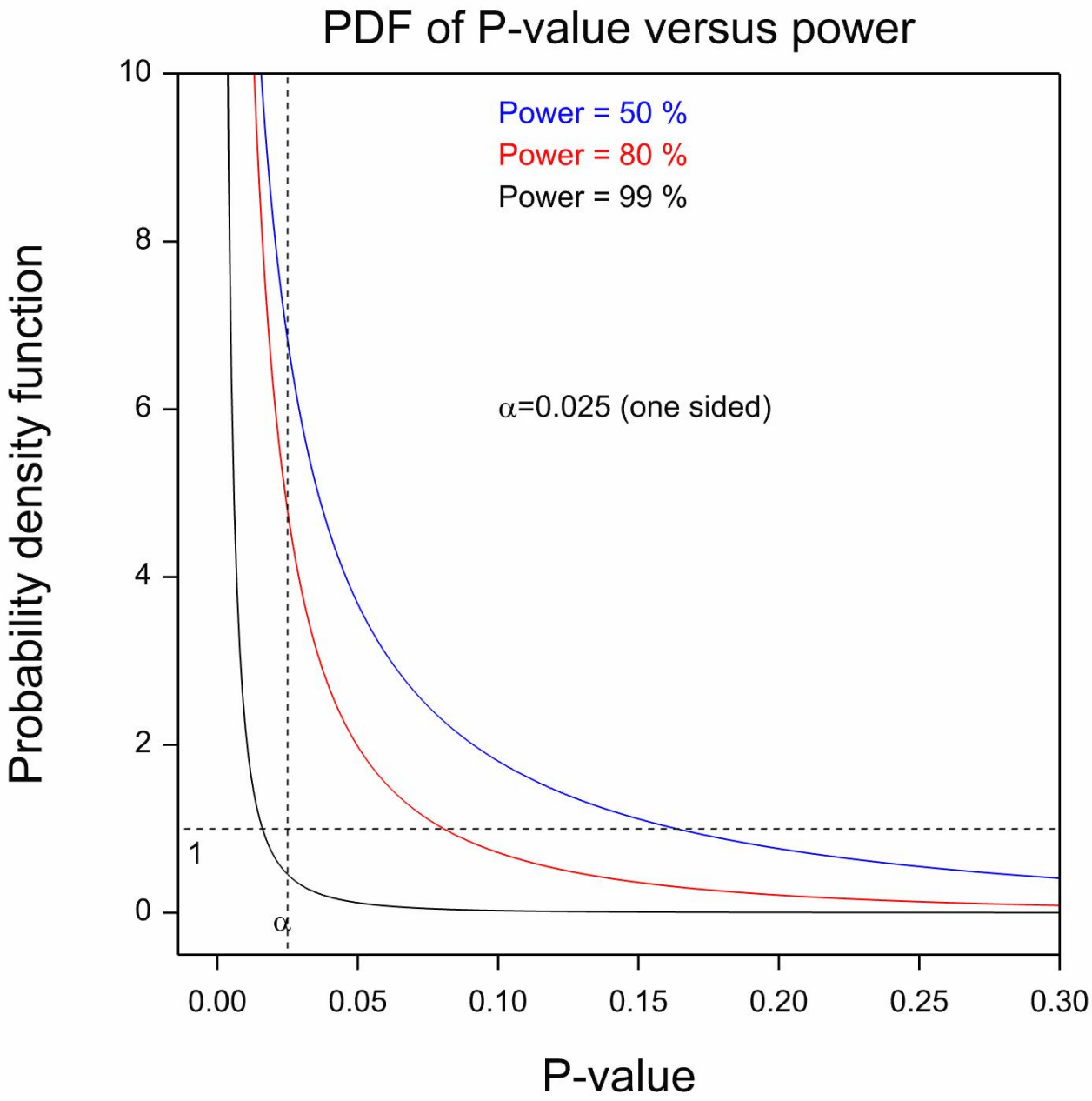
Except when one-sided tests are required by study design, such as in noninferiority trials, all reported P values should be two-sided. In general, P values larger than 0.01 should be reported to two decimal places, those between 0.01 and 0.001 to three decimal places**; P values smaller than 0.001 should be reported as P<0.001**. Notable exceptions to this policy include P values arising in the application of stopping rules to the analysis of clinical trials and genetic-screening studies.

New England Journal of Medicine guidelines to authors

P-values versus significance

- Remember an important distinction
  - significance is $P \leq 0.05$ (say)
  - the P-value might be P=0.047
- If all a Bayesian knows is the former, as power increases the posterior probability of a real effect increases
- On the other hand for the latter as power increases eventually the poster probability decreases
  - Jeffreys-Lindley-Good paradox


True positive rate versus power
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Power (%)
R(1-β)/(α+R(1-β))
α=0.05
R=2
R=1
R=0.5
Based on Ioannidis (2005)


PDF of P-value versus power
Power = 50 %
Power = 80 %
Power = 99 %
α=0.025 (one sided)
Probability density function
P-value
10
8
6
4
2
0
1
α
0.00
0.05
0.10
0.15
0.20
0.25
0.30

A Common Story

- Scientists were treading the path of Bayesian reason
- Along came RA Fisher and persuaded them into a path of P-value madness
- This is responsible for a lot of unrepeatable nonsense
- We need to return them to the path of Bayesian virtue
- In fact the history is not like this and understanding this is a key to understanding the problem

From the table the probability is ·9985 or the odds are about 666 to 1 that 2 is the better soporific.

Student, The Probable Error of a Mean, *Biometrika*, 1908, P21

For $n = 9$, only one value in a hundred will exceed 3·250 by chance, so that the difference between the results is clearly significant.

Fisher, *Statistical Methods for Research Workers*, 1925

The real history

- Scientists before Fisher were using tail area probabilities to calculate posterior probabilities
  - This was following Laplace's use of uninformative prior distributions
- Fisher pointed out that this interpretation was unsafe and offered a *more conservative* one
- Jeffreys, influenced by CD Broad's criticism, was unsatisfied with the Laplacian framework and used a lump prior probability on a point hypothesis being true
  - Etz and Wagenmakers have claimed that Haldane 1932 anticipated Jeffreys
- It is *Bayesian* Jeffreys versus *Bayesian* Laplace that makes the dramatic difference, **not** *frequentist* Fisher versus *Bayesian* Laplace

What Jeffreys Understood

The rule of succession had been generally appealed to as a justification of induction; what Broad showed was that it was no justification whatever for attaching even a moderate probability to a general rule if the possible instances of the rule are many times more numerous than those already investigated. If we are ever to attach a high probability to a general rule, on any practicable amount of evidence, it is necessary that it must have a moderate probability to start with. Thus I may have seen 1 in 1,000 of the 'animals with feathers' in England; on Laplace's theory the probability of the proposition, 'all animals with feathers have beaks', would be about 1/1000. This does not correspond to my state of belief or anybody else's.

Theory of Probability, 3rd edition P128

CD Broad 1887*-1971

- Graduated Cambridge 1910
- Fellow of Trinity 1911
- Lectured at St Andrews & Bristol
- Returned to Cambridge 1926
- Knightbridge Professor of Philosophy 1933-1953
- Interested in epistemology and psychic research

*NB Harold Jeffreys born 1891

CD Broad, 1918

draw counters out of a bag, and, finding that all which we have drawn are white, argue to the probability of the proposition that all in the bag are white.

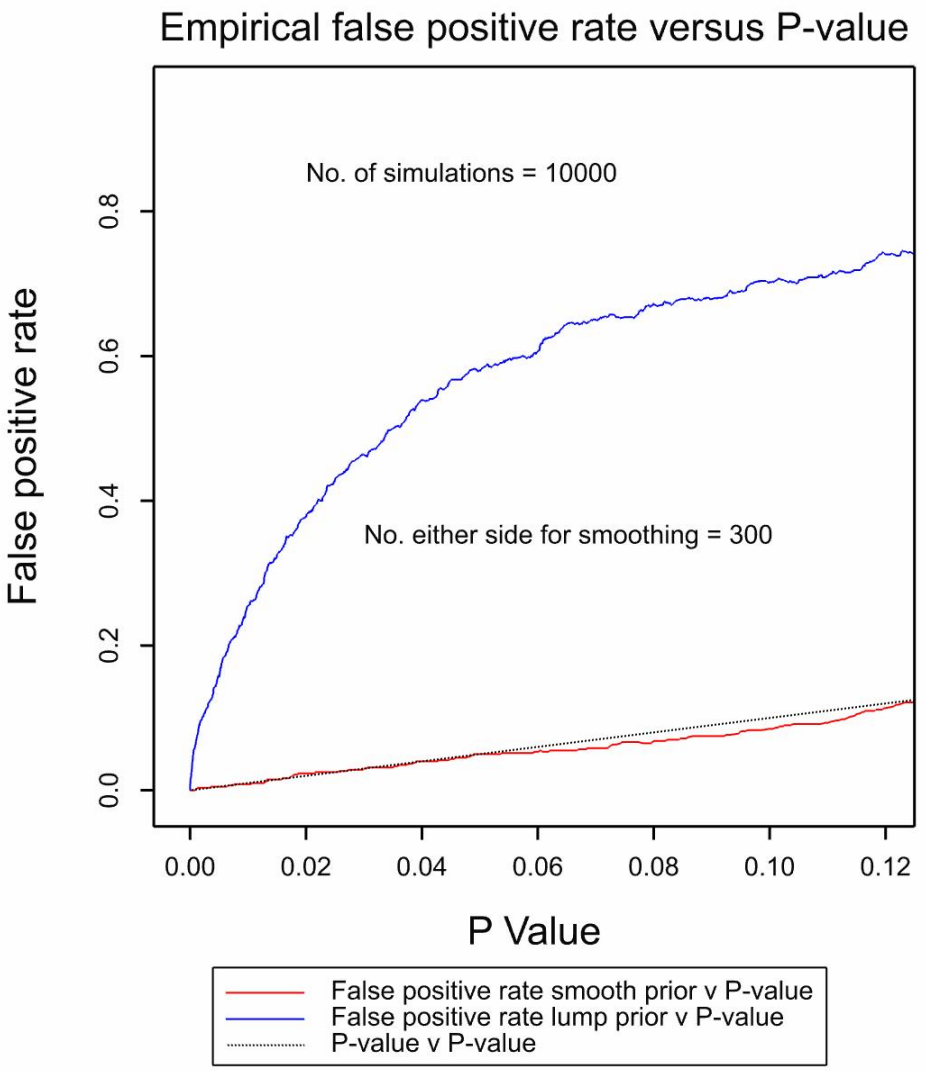
P393

On these assumptions it can be proved that the probability that the *next* to be drawn will be white is $\frac{m+1}{m+2}$, and that the probability that *all* the n are white is $\frac{m+1}{n+1}$.

p394

The Economist gets it wrong

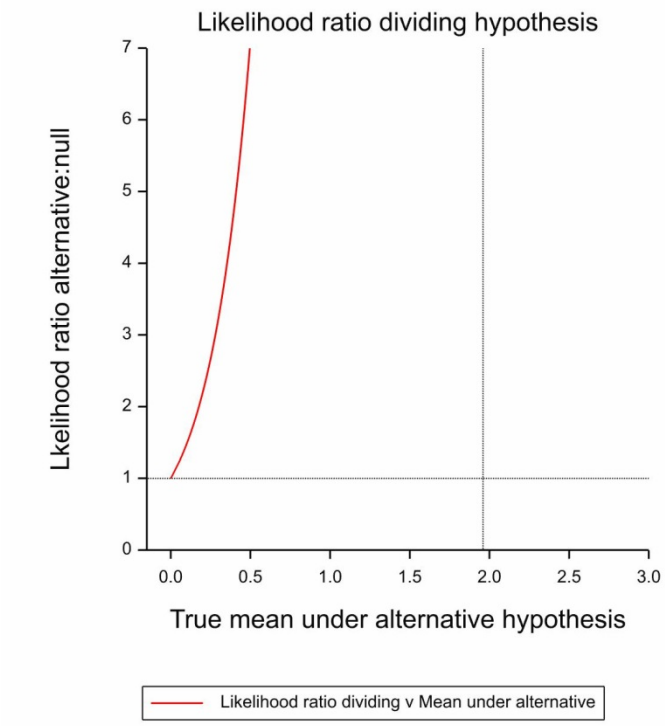
The canonical example is to imagine that a precocious newborn observes his first sunset, and wonders whether the sun will rise again or not. He assigns equal prior probabilities to both possible outcomes, and represents this by placing one white and one black marble into a bag. The following day, when the sun rises, the child places another white marble in the bag. The probability that a marble plucked randomly from the bag will be white (ie, the child's degree of belief in future sunrises) has thus gone from a half to two-thirds. After sunrise the next day, the child adds another white marble, and the probability (and thus the degree of belief) goes from two-thirds to three-quarters. And so on. Gradually, the initial belief that the sun is just as likely as not to rise each morning is modified to become ***a near-certainty that the sun will always rise***.

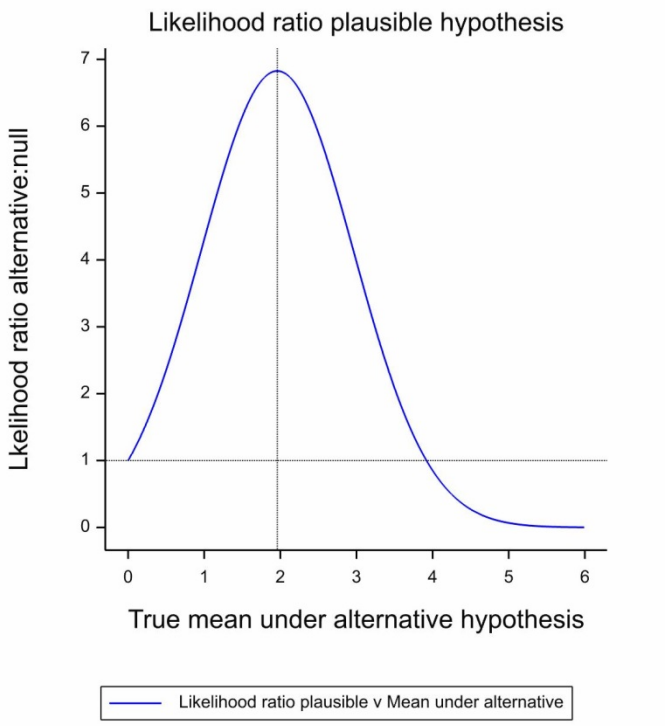

Empirical false positive rate versus P-value
No. of simulations = 10000
No. either side for smoothing = 300
False positive rate
0.0
0.2
0.4
0.6
0.8
0.00
0.02
0.04
0.06
0.08
0.10
0.12
P Value
False positive rate smooth prior v P-value
False positive rate lump prior v P-value
P-value v P-value

Why the difference?

- Imagine a point estimate of two standard errors
- Now consider the likelihood ratio for a given value of the parameter, δ under the alternative to one under the null
  - *Dividing hypothesis (smooth prior)* for any given value $\delta = \delta'$ compare to $\delta = -\delta'$
  - Plausible hypothesis (lump prior) for any given value $\delta = \delta'$ compare to $\delta = 0$

The situations compared





A speculation of mine

- Scientists had noticed that for dividing hypotheses they could get 'significance' rather easily
  - The result is the 1/20 rule
- However when deciding to choose a new parameter or not in terms of probability it is 50% not 5% that is relevant
- This explains the baffling finding that significance test are actually *more* conservative than AIC (and sometimes than BIC)

The illogicality of asking for 95% posterior probability

- If we are going to switch from significance to posterior probabilities we need to recalibrate
- 1/20 belongs to the significance system
- It is not appropriate for the posterior probability system
- To switch & keep would be like abandoning dosing by age for dosing by bodyweight & saying that because you had to be aged 10 to take the medicine you had to be at least 10kg in weight to take it

Goodman's Criticism

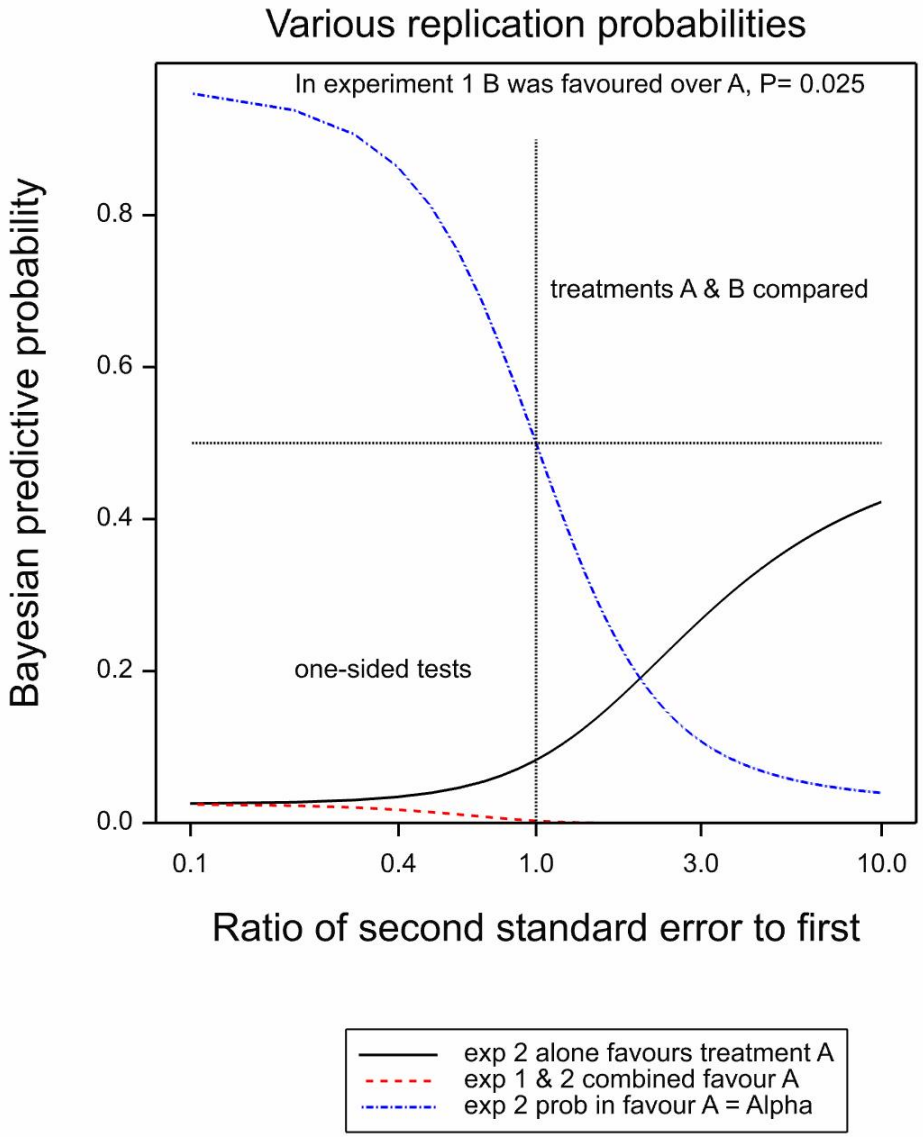
- What is the probability of repeating a result that is just significant at the 5% level (p=0.05)?
- Answer 50%
  - If true difference is observed difference
  - If uninformative prior for true treatment effect
- Therefore P-values are unreliable as inferential aids

Sauce for the Goose and 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

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?


Various replication probabilities
In experiment 1 B was favoured over A, P= 0.025
treatments A & B compared
one-sided tests
Bayesian predictive probability
0.0
0.2
0.4
0.6
0.8
0.1
0.4
1.0
3.0
10.0
Ratio of second standard error to first
exp 2 alone favours treatment A
exp 1 & 2 combined favour A
exp 2 prob in favour A = Alpha

Why Goodman's Criticism is Irrelevant

"It would be absurd if our inferences about the world, having just completed a clinical trial, were *necessarily* dependent on assuming the following. 1. We are now going to repeat this experiment. 2. We are going to repeat it only once. 3. It must be exactly the same size as the experiment we have just run. 4. The inferential meaning of the experiment we have just run is the extent to which it predicts this second experiment."

Senn, 2002

A Paradox of Bayesian Significance Tests

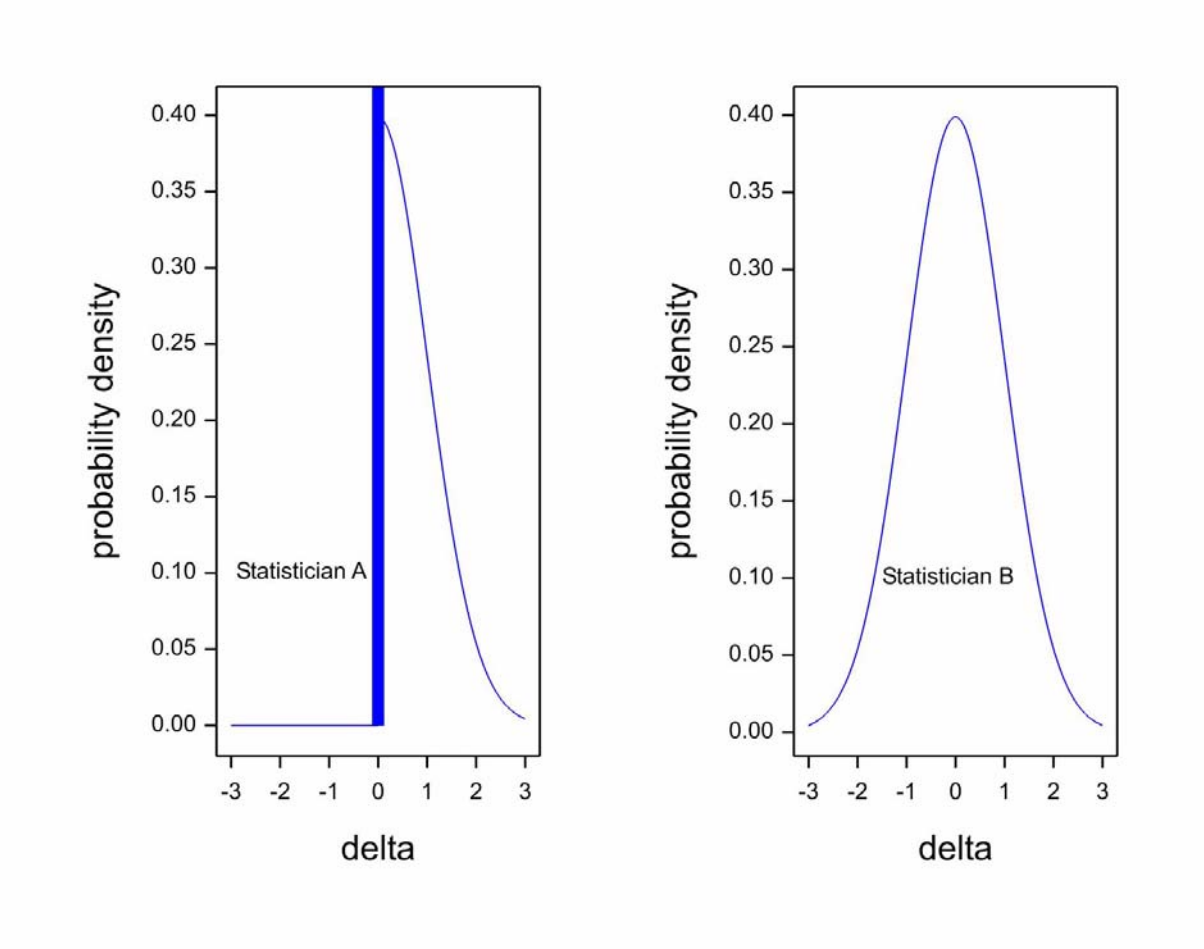
Two scientists start with the same probability 0.5 that a drug is effective.

Given that it is effective they have the same prior for how effective it is.

If it is not effective A believes that it will be useless but B believes that it may be harmful.

Having seem the same data B now believes that it is useful with probability 0.95 and A believes that it is useless with probability 0.95.

A Tale of Two priors



In Other Words

The probability is 0.95

And the probability is also 0.05

Both of these probabilities can be simultaneously true.

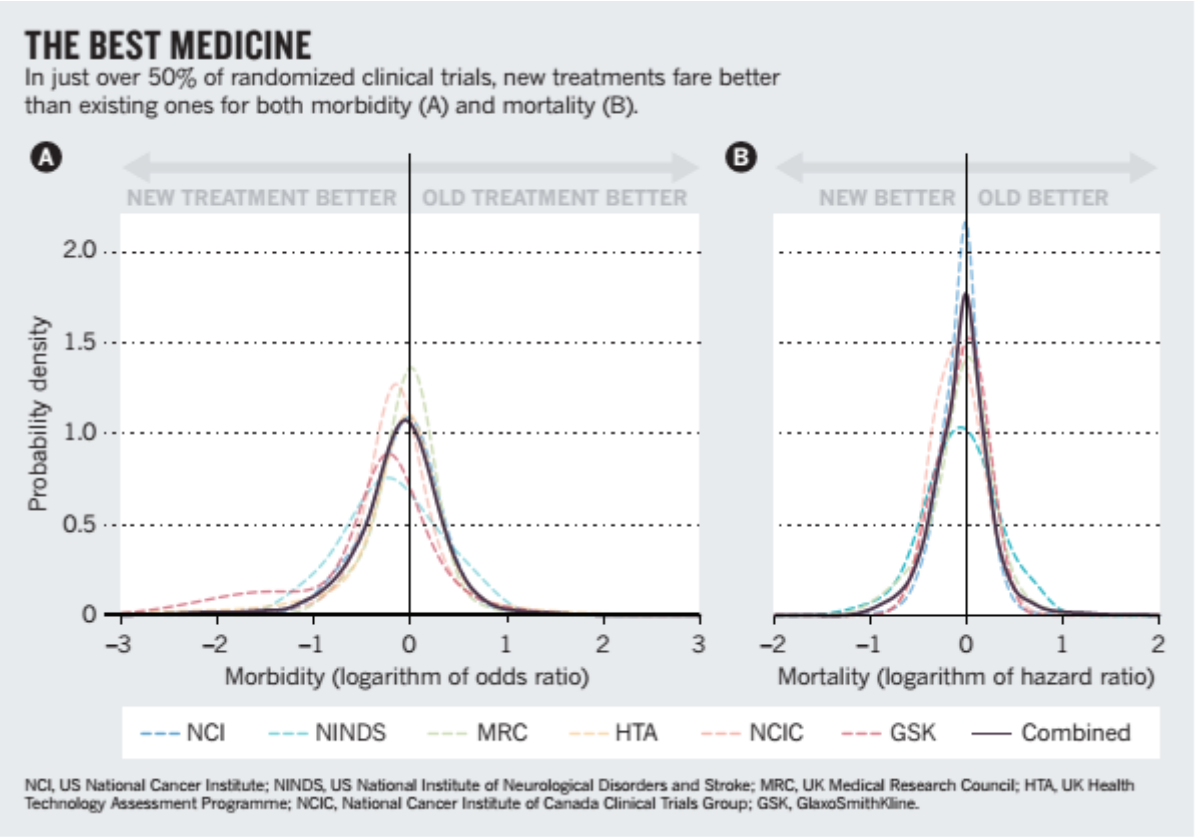
NB This is *not* illogical but it is illogical to regard this sort of thing as proving that p-values are illogical

'…would require that a procedure is dismissed because, when combined with information which it doesn't require and which may not exist, it disagrees with a procedure that disagrees with itself'

Senn, 2001, Two cheers for P-values

What do scientists really test?

- David Colquhoun maintains that the fact that scientists use two-sided tests means they are testing $H_1: \tau \neq 0 \; v \; H_0: \tau = 0$
  - But to claim that two sided tests imply a point null is to confuse a sufficient condition for a necessary one
  - Two one sided null hypotheses with split alpha is another possibility
    - $H_{1a}: \tau > 0 \; v \; H_{0a}: \tau \leq 0$, $H_{1b}: \tau < 0 \; v \; H_{0b}: \geq 0$
  - And it could be argued that in drug development so-called two-sided tests at the 5% level are really one-sided at the 2.5% level
- But in any case it is not what scientists believe that is relevant to false positive rates but what nature determines

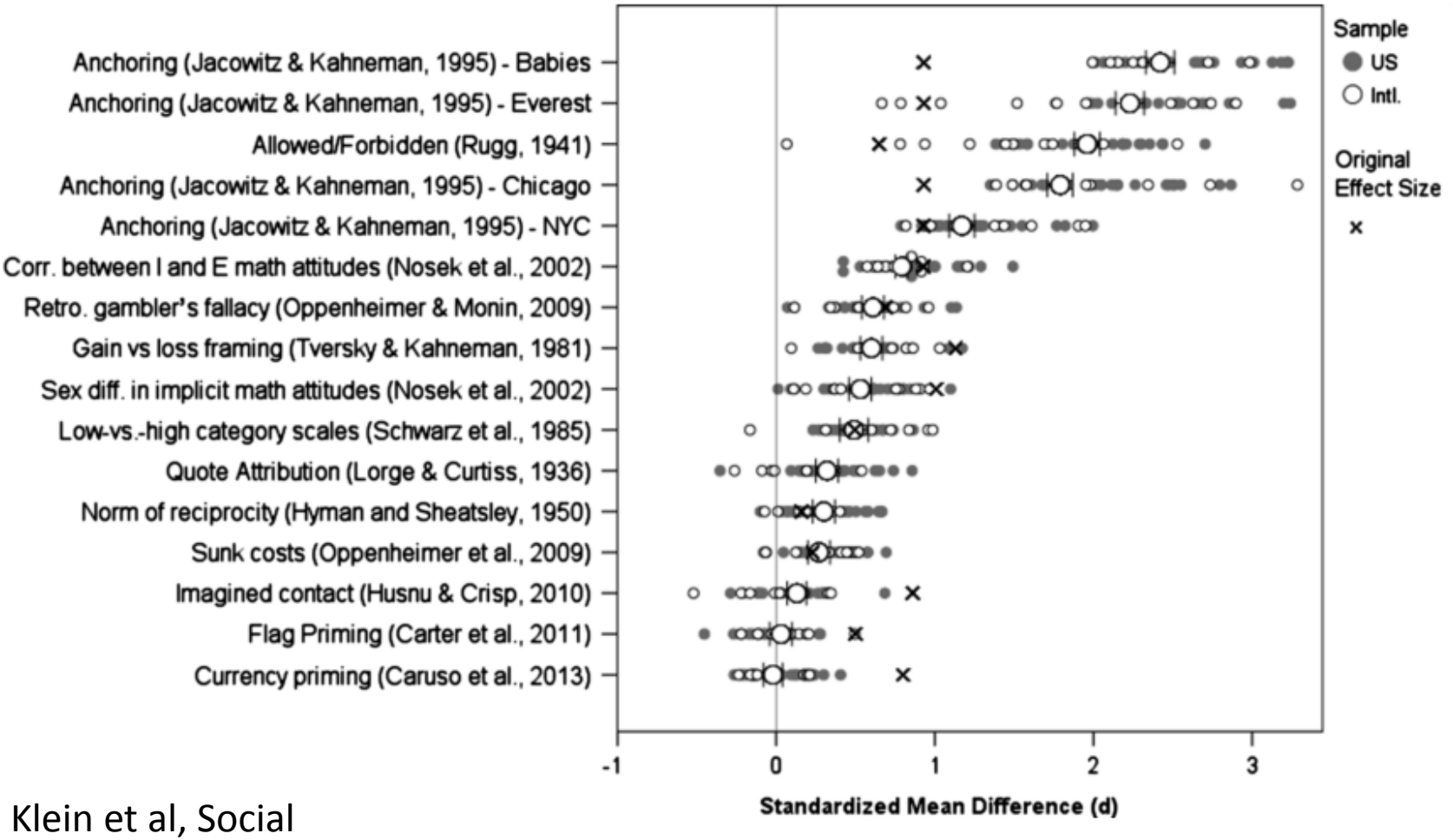


860 trials in 350,000 patients reported by Djulbegovic et al, Nature, August 2013

Are most research findings false?

- A dram of data is worth a pint of pontification
- Two interesting studies recently
  - The many labs replications project
    - This raised the Twitter storm alluded to earlier
  - Jager & Leek, *Biostatistics* 2014

Many Labs Replication Project



Klein et al, Social Psychology, 2014

Jager & Leek, 2014

- Text-mining of 77,410 abstracts yielded 5,322 P-values
- Considered a mixture model truncated at 0.05
- Estimated that amongst 'discoveries' 14% are false

$$f(p|a,b,\pi_0) = \pi_0 uniform(0,0.05) + (1-\pi_0) tBeta(a,b,0.05)$$

Estimation using the EM algorithm

But one must be careful

- These studies suggest that a common threshold of 5% seems to be associated with a manageable false positive rate
- This does not mean that the threshold is right
  - It might reflect (say) that most P-values are either > 0.05 or << 0.05
    - Remember the P-value PDF
  - The situation might be capable of improvement using a different threshold
- Also, are false negatives without cost?

My Conclusion

- 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

In defense of P values

PAUL A. MURTAUGH

Abstract. Statistical hypothesis testing has been widely criticized by ecologists in recent years. I review some of the more persistent criticisms of P values and argue that most stem from misunderstandings or incorrect interpretations, rather than from intrinsic shortcomings of the P value**. I show that P values are intimately linked to confidence intervals and to differences in Akaike's information criterion (DAIC)**, two metrics that have been advocated as replacements for the P value. The choice of a threshold value of DAIC that breaks ties among competing models is as arbitrary as the choice of the probability of a Type I error in hypothesis testing, and **several other criticisms of the P value apply equally to DAIC**. Since P values, confidence intervals, and DAIC are based on the same statistical information, all have their places in modern statistical practice. **The choice of which to use should be stylistic, dictated by details of the application rather than by dogmatic, a priori considerations.**

Ecology, 95(3), 2014, pp. 611–617

Wise words

Proponents of the “Bayesian revolution” should be wary of chasing yet another chimera: an apparently universal inference procedure. A better path would be to promote both an understanding of the various devices in the “statistical toolbox” and informed judgment to select among these.

Gigerenzer and Marewski,
Journal of Management, Vol. 41 No. 2, February 2015 421–440