P Value wars

Stephen Senn

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Acknowledgements

Acknowledgements

Thanks to the EMA for inviting me and to Olivier Collignon 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

- Basics
  - The difference between probability and statistics
  - A simple example
- Recent criticisms of P-values/significance
  - Ioannides
  - Repligate and other psychology fights
  - ASA statement
- A brief history of P-values
  - The real (hidden) reason why P-values are so controversial
- Conclusions

P-values, Bayes The meaning of life

BASICS

Probability versus Statistics an example

Probability theory

- Q. This die is fair, what is the probability that I will get two sixes in two rolls of the die
- A. $\left(\frac{1}{6}\right)^2 = \frac{1}{36}$

Statistical theory

- Q I rolled this die twice and got two sixes. What is the probability that the dice is fair?
- A. Nobody knows

Probabilists versus Statisticians: the difference

Probabilists

- Are mathematicians
- Deal with direct probability statements
- Plays a formal mathematical game
- Are involved in the divine
  - God knows that dice is fair; let's work out the consequences

Statisticians

- Are scientists
- Deal with inverse probability statements
- Are dealing with the real world
- Are involved in the human
  - We are down here trying to work out the mind of God

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

That Heavy Metal Problem
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

What is the difference between the Bayesian probability and the P-value?

Bayesian

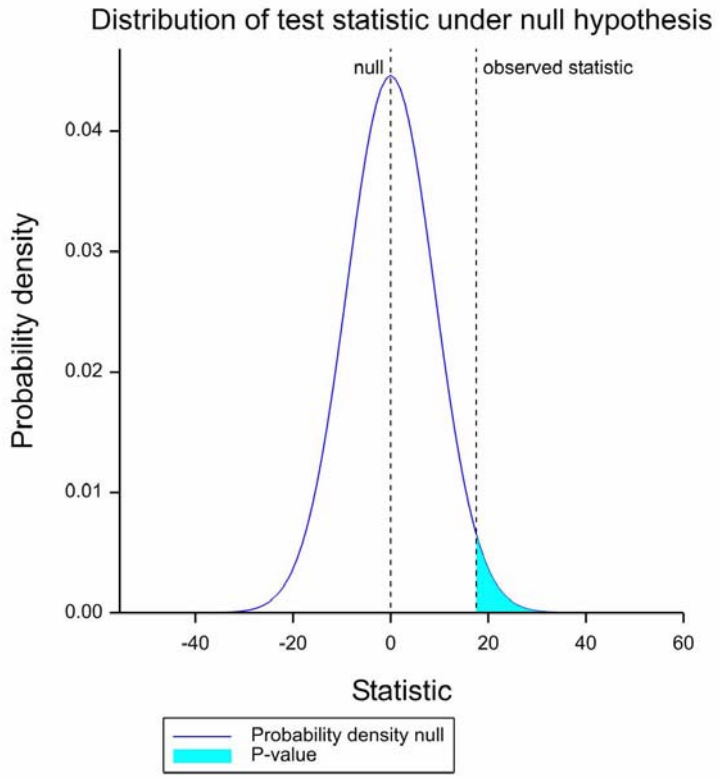
- It is the probability that the null hypothesis is true given the evidence
  - It is my posterior probability that I pressed shuffle
- The value is 9/21

P-value

- It is the probability of the evidence given the null hypothesis
  - Usually more extreme evidence must be included
- The value is 1/12

Extremism

- For many practical applications *every* result is unlikely
- Example: the probability that the mean difference in blood pressure is exactly 5mmHg is effectively zero
- This has led to calculation of probability of result *as extreme or more extreme* to act as ‘p-value’



NB Probability statements are not reversible

- Is the Pope a Catholic?
  - Yes
- Is a Catholic the Pope?
  - Probably not
- The probability of the evidence given the hypothesis is not the same as the probability of the hypothesis given the evidence
  - As we saw from the CD player example

To sum up

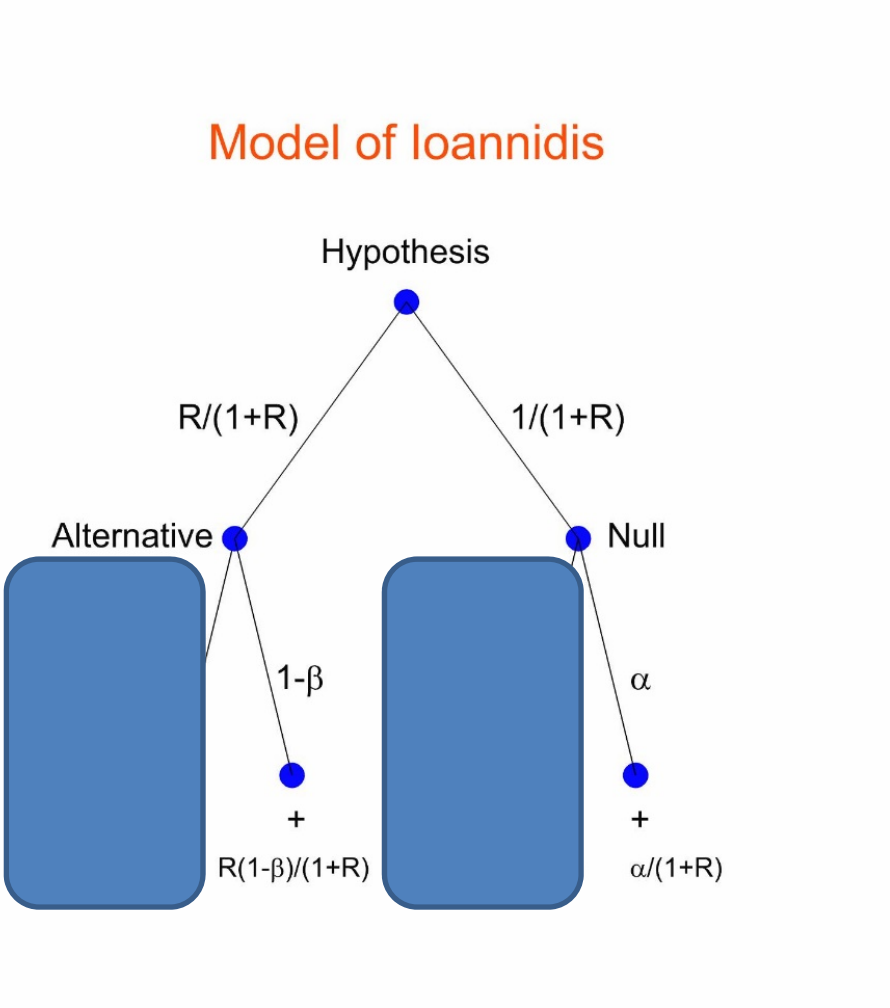
- The Bayesian approach provides a complete and logical solution
- But it requires prior probabilities
- You can’t just use arguments of symmetry
- Different people will come to different conclusions

Ioannides, Repligate and the ASA statement

RECENT CRITICISMS

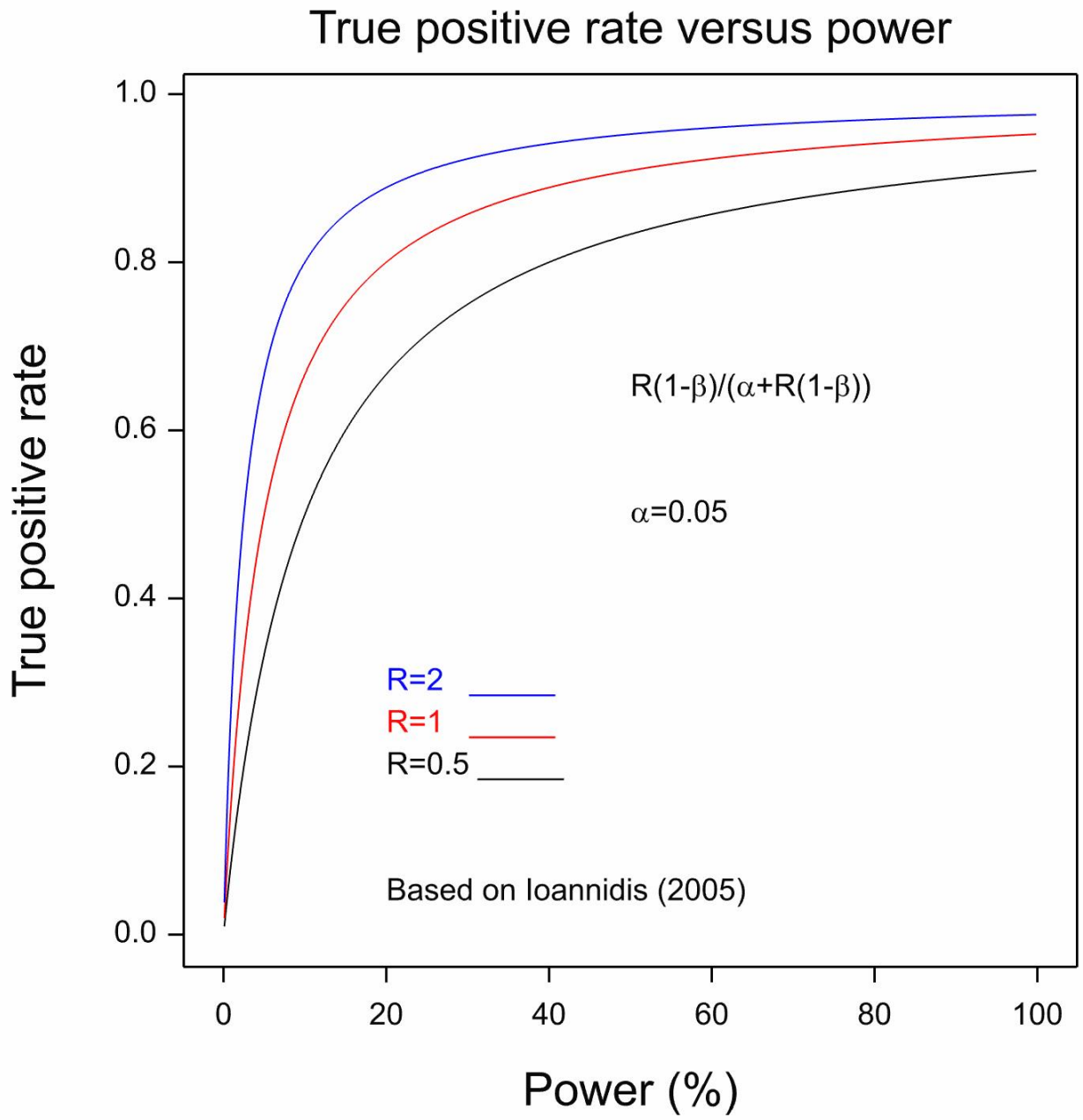
Ioannidis (2005)

- Claimed that most published research findings are wrong
  - By finding he means a 'positive' result
- 4380 citations by 16 February 2017 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*)


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)

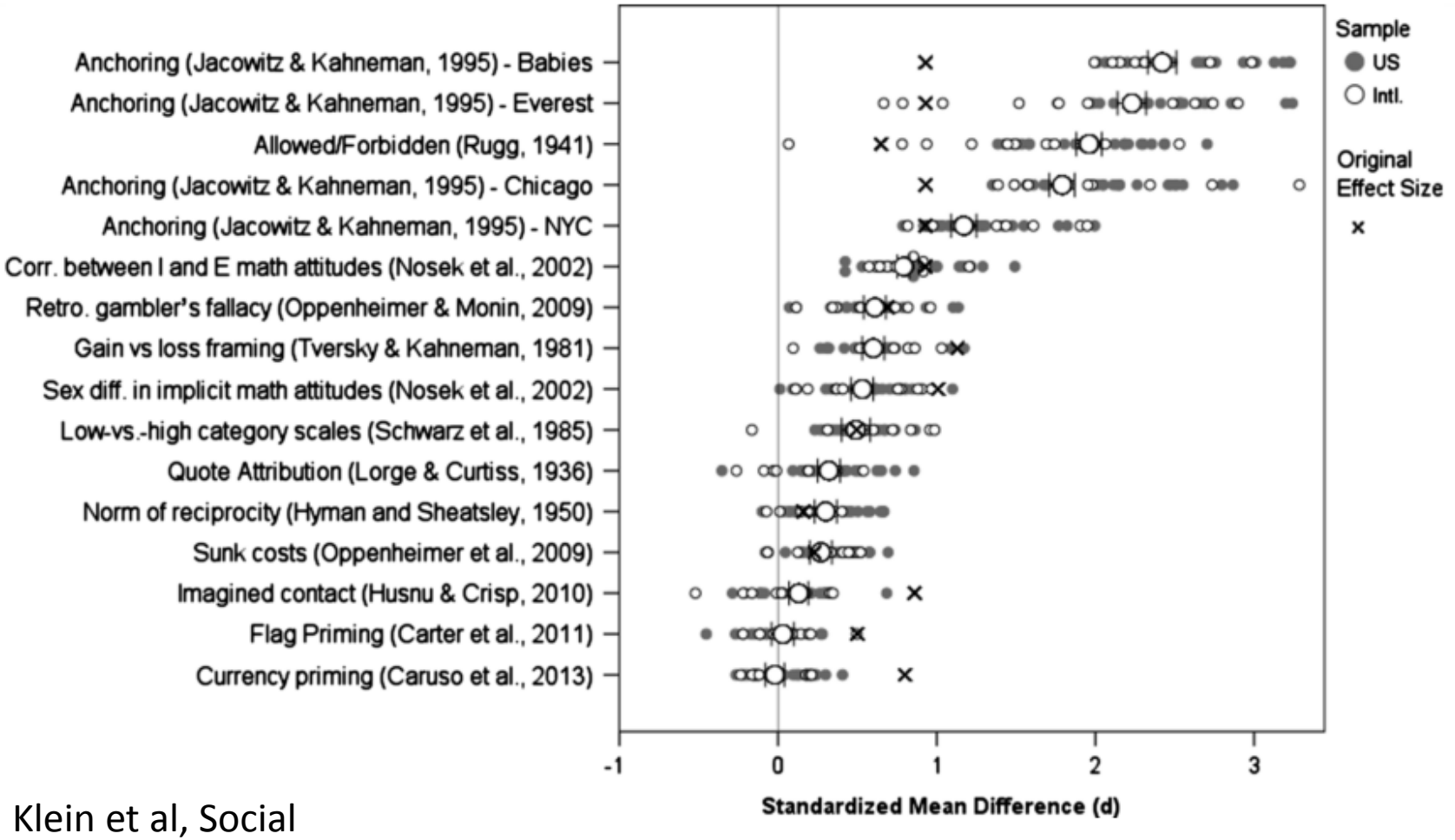
The Crisis of Replication

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*

Many Labs Replication Project



Klein et al, Social Psychology, 2014

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research [1].

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Auth

Archive > Volume 531 > Issue 7593 > News > Article

NATURE | NEWS

Statisticians issue warning over misuse of *P* values

Policy statement aims to halt missteps in the quest for certainty.

Monya Baker

07 March 2016

PDF | Rights & Permissions

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warns in a statement released today[1]. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied in ways that cast doubt on statistics generally, he adds.

The ASA statement

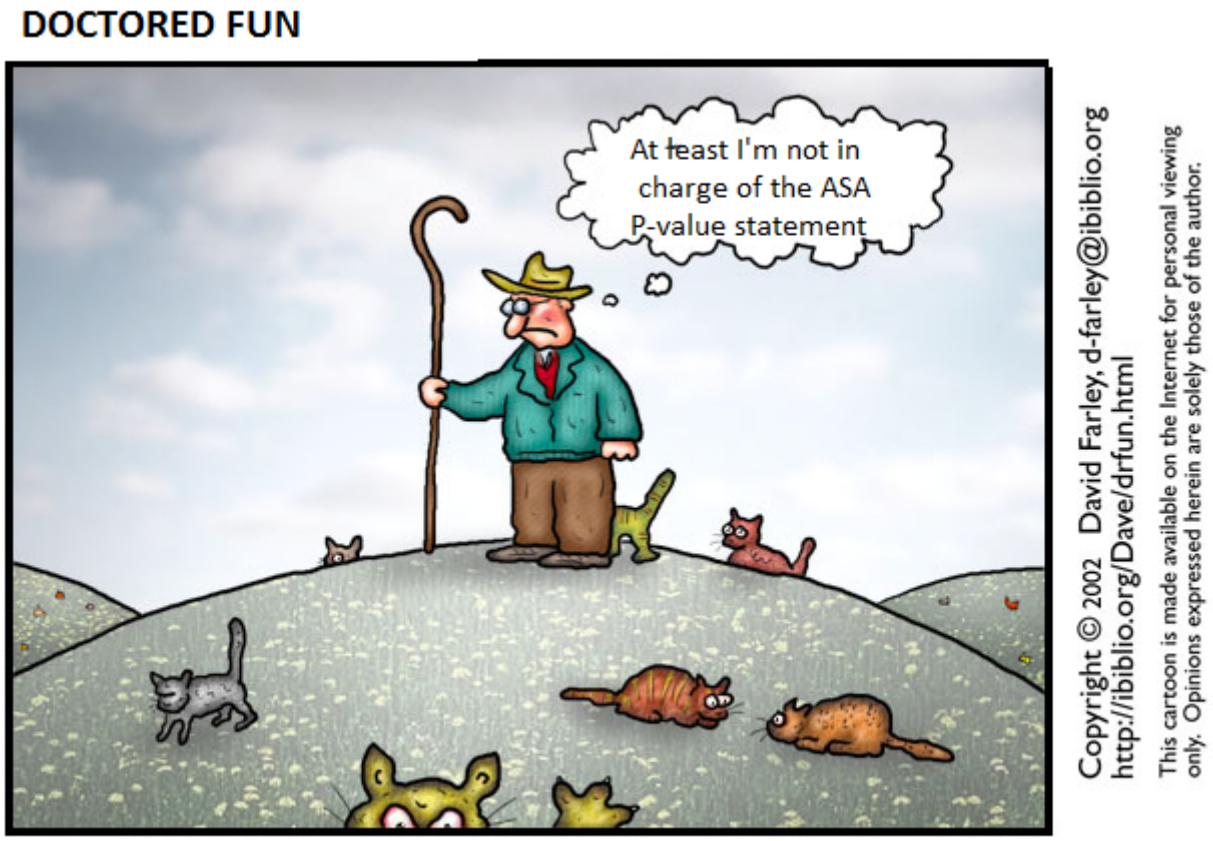
1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Obligatory purloined cartoon

The daydreams of cat herders

(c) Stephen Senn
2017

Sympathy for Ron Waserstein



The consolation of cat-herders

In summary

- There has been mounting criticism of ‘null-hypothesis-significance-testing’ (NHST) and P-values in particular
- Some people think that ‘science is broke’
- Some people claim that P-values are to blame
  - They give ‘significance’ far too easily
- There is a lot of advice, much of it conflicting, flying around

An Example of the Problem

"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."

David Colquhoun

Royal Society Open Science 2014

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**

New England Journal of Medicine guidelines to authors

Did Fisher really teach scientists to fish for significance?

A BRIEF HISTORY OF P-VALUES

The collective noun for statisticians is “a quarrel”

John Tukey

A Common Story

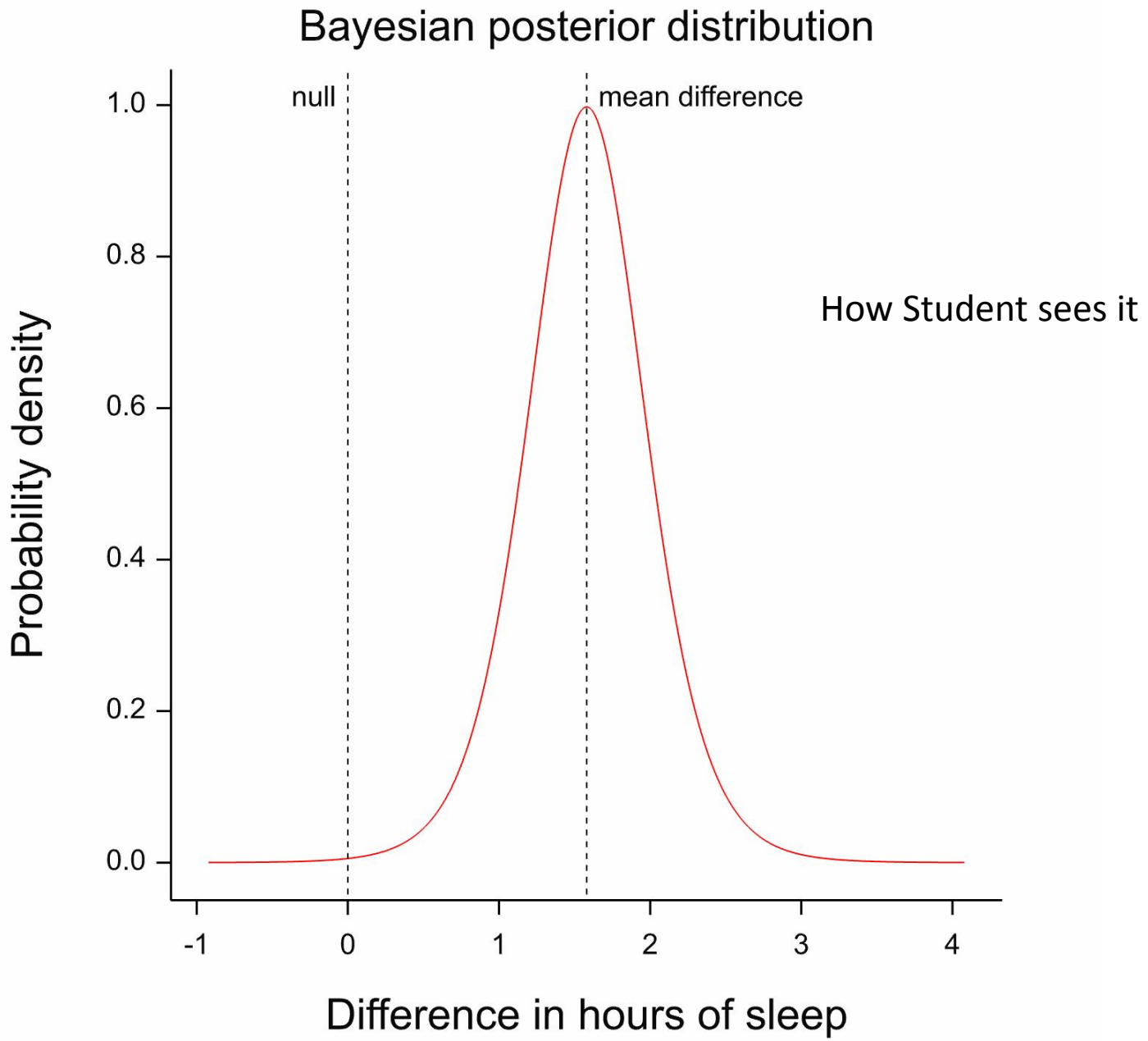
- Scientists were treading the path of Bayesian reason
- Along came RA Fisher and persuaded them into a path of P-value madness
- This is responsible for a lot of unrepeatable nonsense
- We need to return them to the path of Bayesian virtue
- In fact the history is not like this and understanding this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

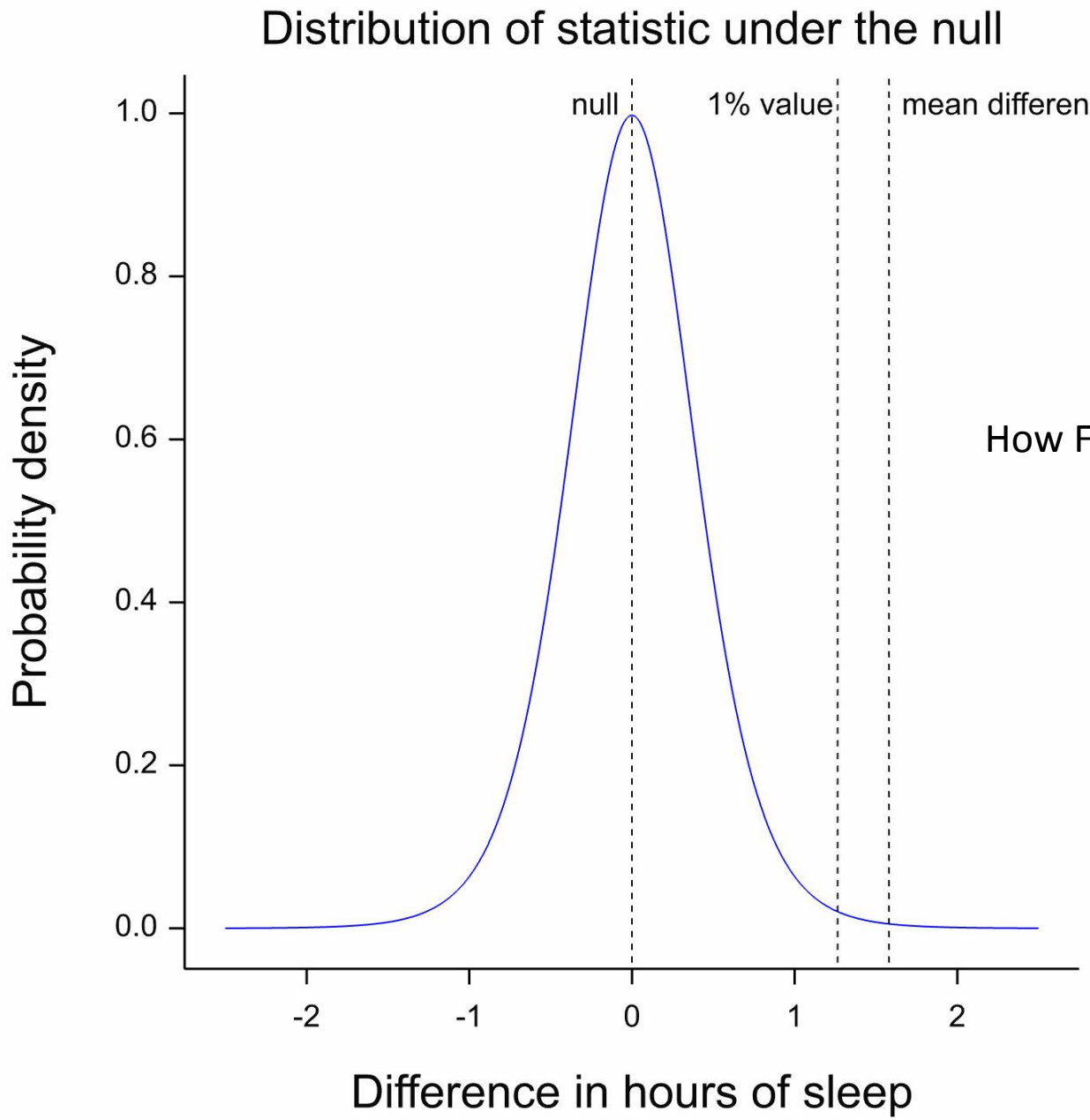
Student, The Probable Error of a Mean, *Biometrika*, 1908, P21

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.

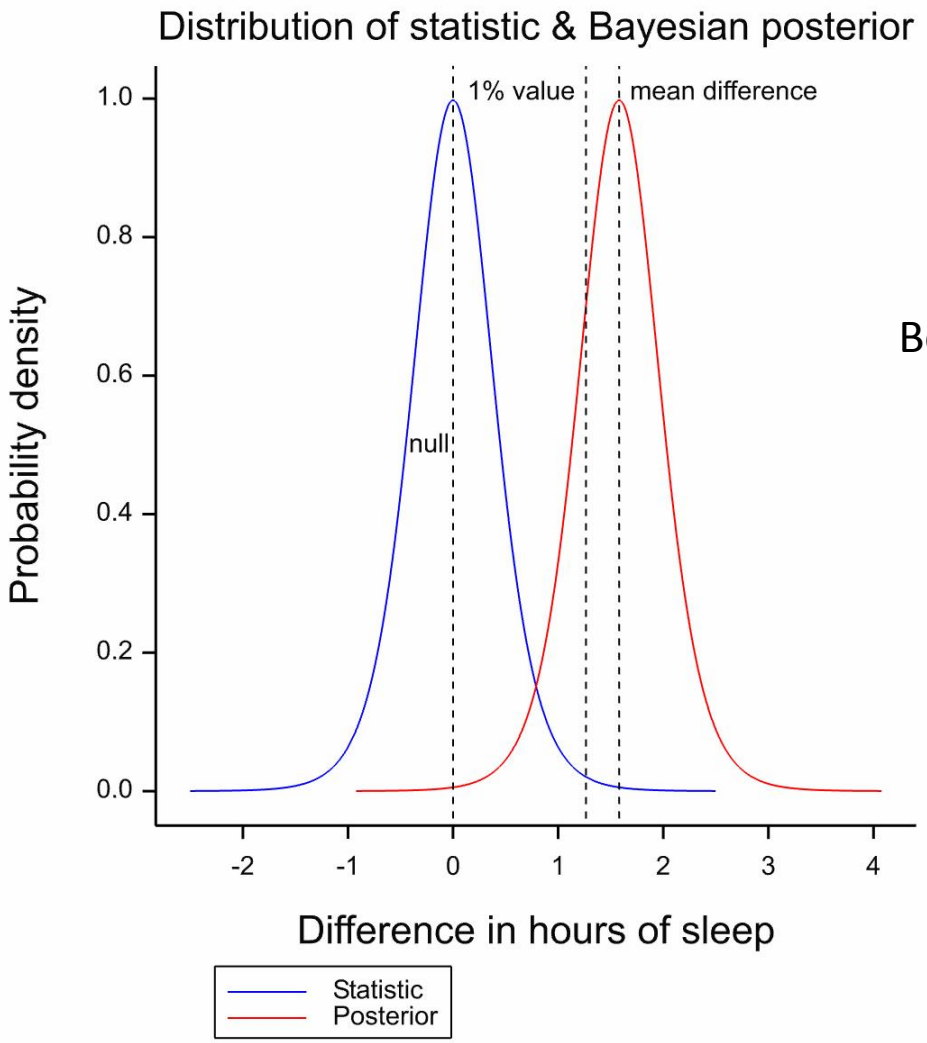
Fisher, *Statistical Methods for Research Workers*, 1925



How Student sees it



How Fisher sees its



Both points of view

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.

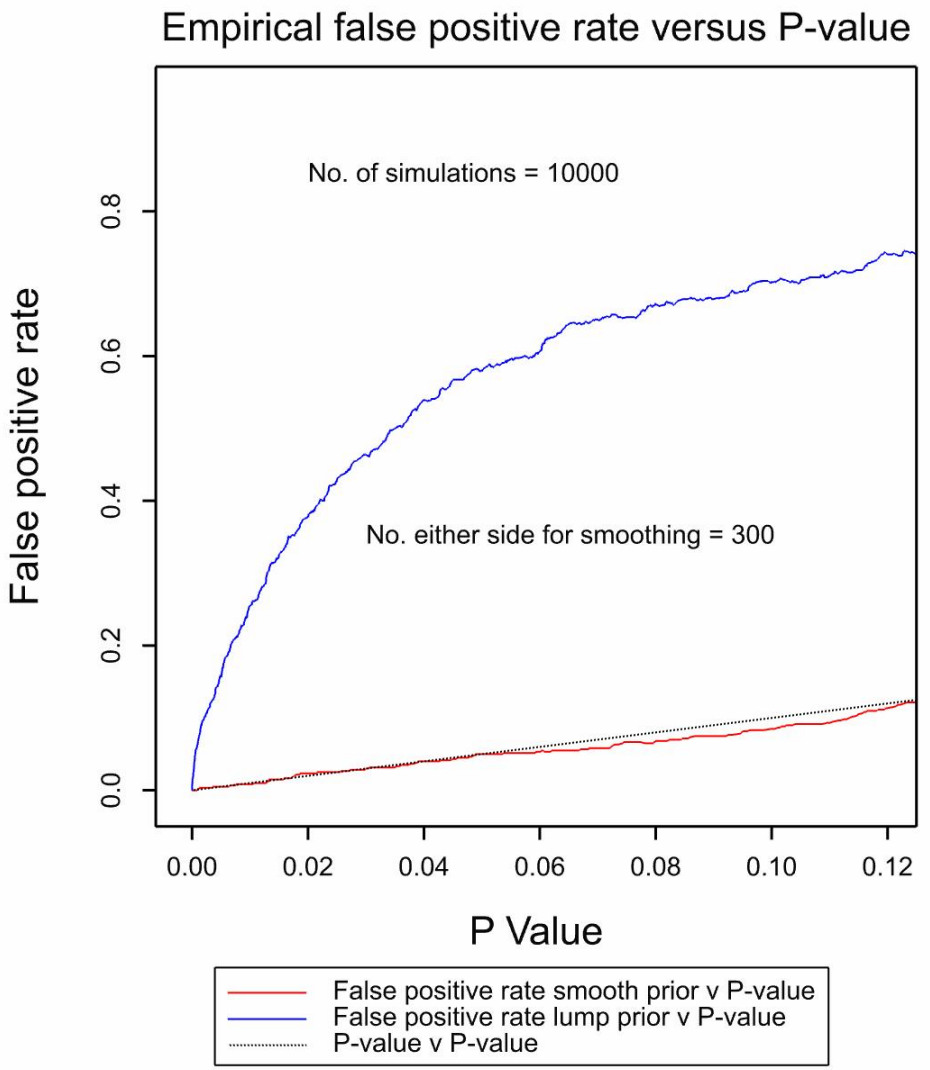
P393

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

What Jeffreys concluded

- If you have an uninformative prior distribution the probability of a precise hypothesis is very low
- It will remain low even if you have lots of data consistent with it
- You need to allocate a solid lump of probability that it is true
- Nature has decided, other things being equal, that simpler hypotheses are more likely


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 $\delta = 0$

In summary

- The major disagreement is not between P-values and Bayes using informative prior distribution
- It's between two Bayesian approaches
  - Using uninformative prior distributions
  - Using highly informative one
- The conflict is not going to go away by banning P-values
- There is no automatic Bayesianism
  - You have to do it for real

Statistics is difficult and statisticians should be paid more

CONCLUSION

What you should know

- P=0.05 is a weak standard of evidence
- Requiring two trials is a good idea
  - NB If you replace them by a single larger trial or a meta-analysis ask for P=1/800
  - Variation in results from trial to trial is natural
- Don't just rely on P-values
- Bayes is harder to do than many people think
- Pay attention to the ASA advice

In Conclusion
The Solid Six?

1. P-values can indicate how incompatible the data are with a specified statistical model. ✔ ✔
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. ✔ ✔ ✔
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. ✔ ✔ ✔
4. Proper inference requires full reporting and transparency. ✔ ✔ ✔
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. ✔ ✔ ✔
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. ✔ ✔

However

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