

# P Value wars

Stephen Senn



# Acknowledgements

## Acknowledgements

Thanks to the Richard Morey and the Doctoral Academy for inviting me and to David Mehler 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
  - An example
- Recent criticisms of P-values/significance
  - Ioannides
  - Repligate
  - 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**

# Probabilists versus Statisticians: the difference

## Probabilists

- Are mathematicians
- Deal with direct probability statements
- Play a formal mathematical game
- Are involved in the divine
  - God knows that dice are 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

# 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 die is fair?
- A. Nobody knows

# 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

# Two Solutions

## Bayesian

- State a prior probability for every possible degree of bias in the die
- Calculate the likelihood (the probability of the observed result as a function of every possible degree of bias)
- These two via Bayes theorem yield a posterior probability

## Frequentist

- Assume the die is fair
- Calculate the probability of the result on the assumption that the die is fair
- If this probability is low, cast doubt on the fairness of the die



# 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

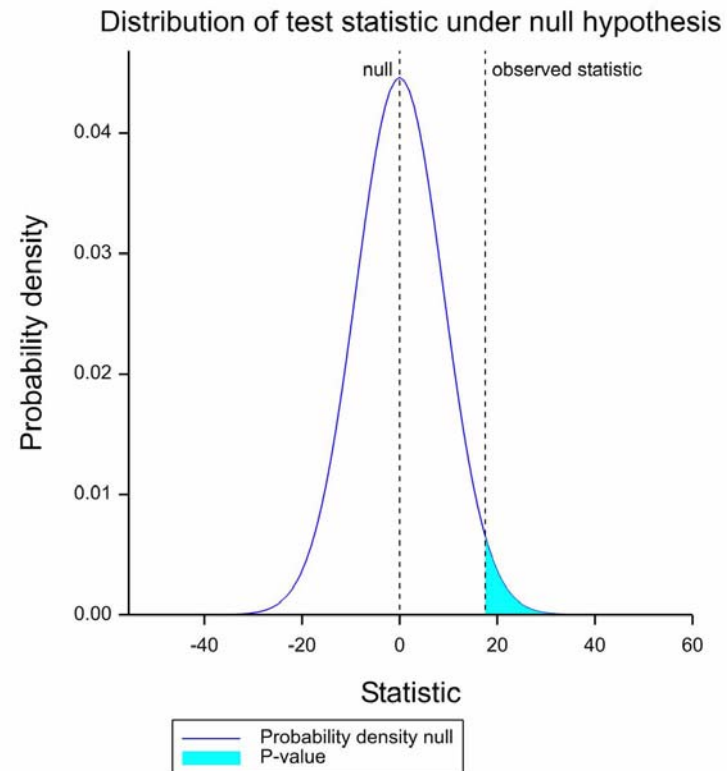
Hypothesis	Prior Probability P	Evidence	Likelihood	P x L
Shuffle	9/10	W	1/12	9/120
Play	1/10	W	1	12/120
TOTAL				21/120

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

# 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'



# 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

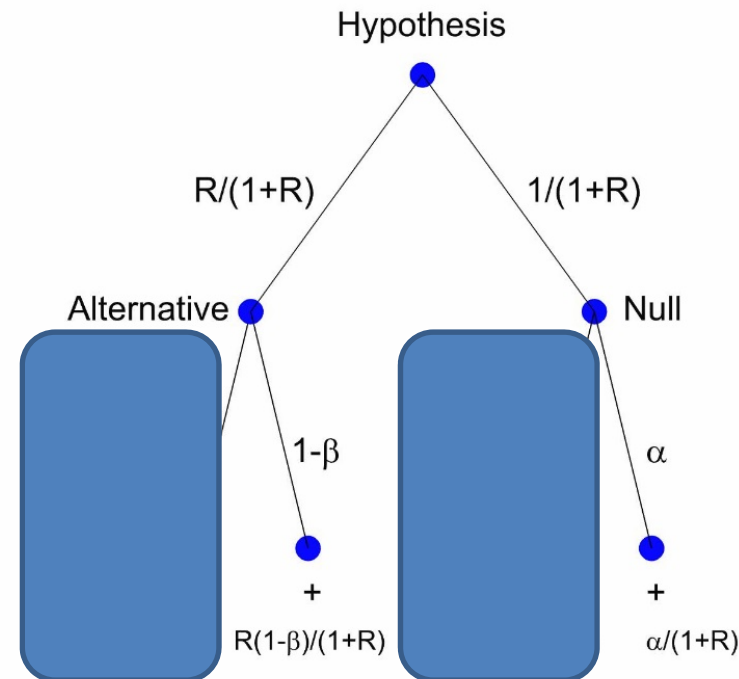
# Ioannidis (2005)

- Claimed that most published research findings are wrong
  - By finding he means a ‘positive’ result
- 4491 citations by 26 March 2017 according to Google Scholar

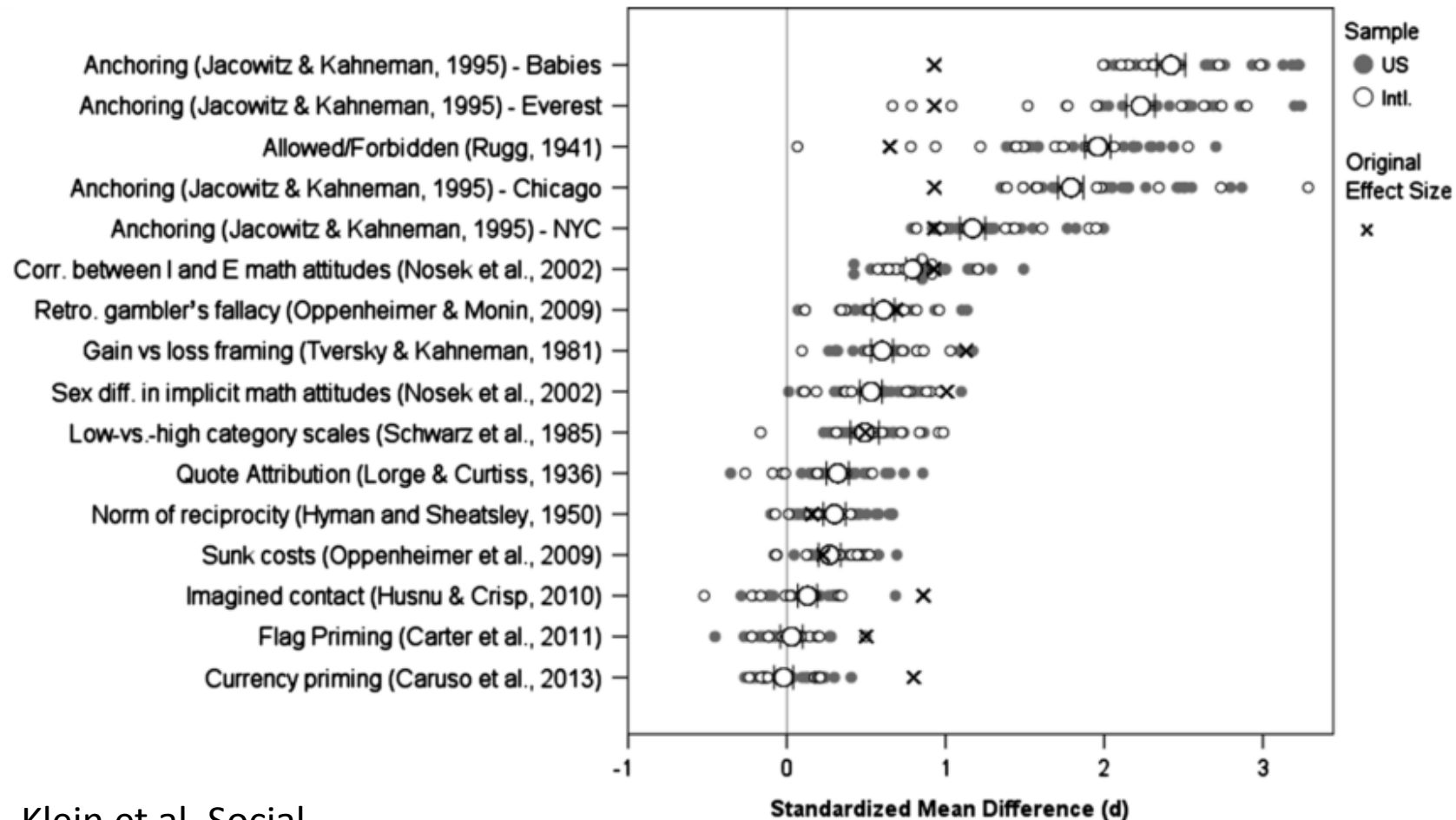
$$TPR = \frac{R(1-\beta)/(1+R)}{[\alpha+R(1-\beta)]/(1+R)} = \frac{R(1-\beta)}{\alpha+R(1-\beta)}$$

(TPR= True Positive Rate)

## Model of Ioannidis



# Many Labs Replication Project



Klein et al, Social Psychology, 2014

(c) Stephen Senn 2017

# 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

## DOCTOR FUN

| Oct 2002



Copyright © 2002 David Farley, d-farley@ibiblio.org  
<http://ibiblio.org/Dave/drfun.html>

This cartoon is made available on the Internet for personal viewing only. Opinions expressed herein are solely those of the author.

The daydreams of cat herders

(c) Stephen Senn  
2017

# Sympathy for Ron Wasserstein

DOCTORED FUN



Copyright © 2002 David Farley, d-farley@ibiblio.org  
<http://ibiblio.org/Dave/drfun.html>

This cartoon is made available on the Internet for personal viewing only. Opinions expressed herein are solely those of the author.

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

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.

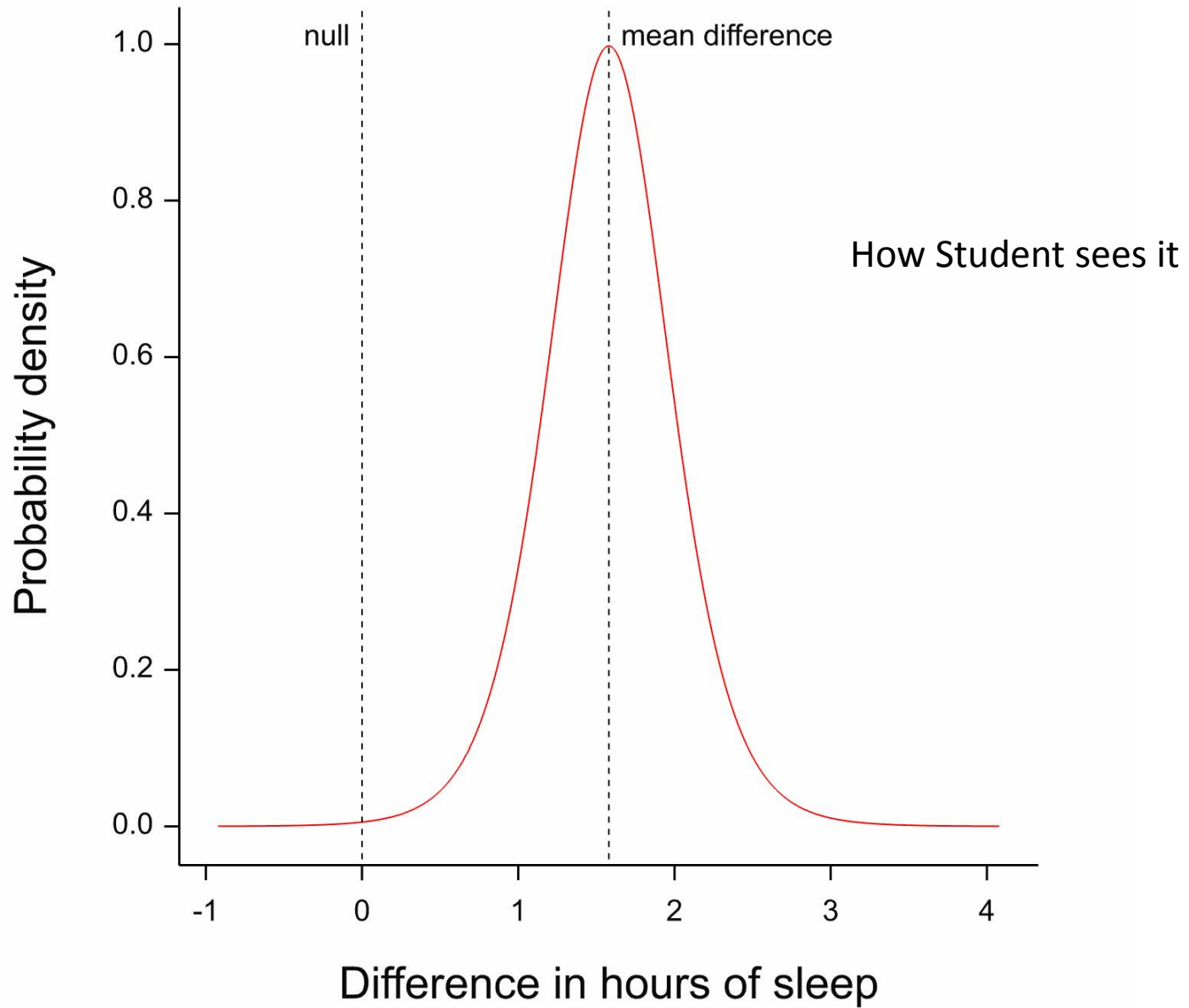
Student, The Probable Error of a Mean, *Biometrika*, 1908, P21

122 STATISTICAL METHODS [§ 24·1

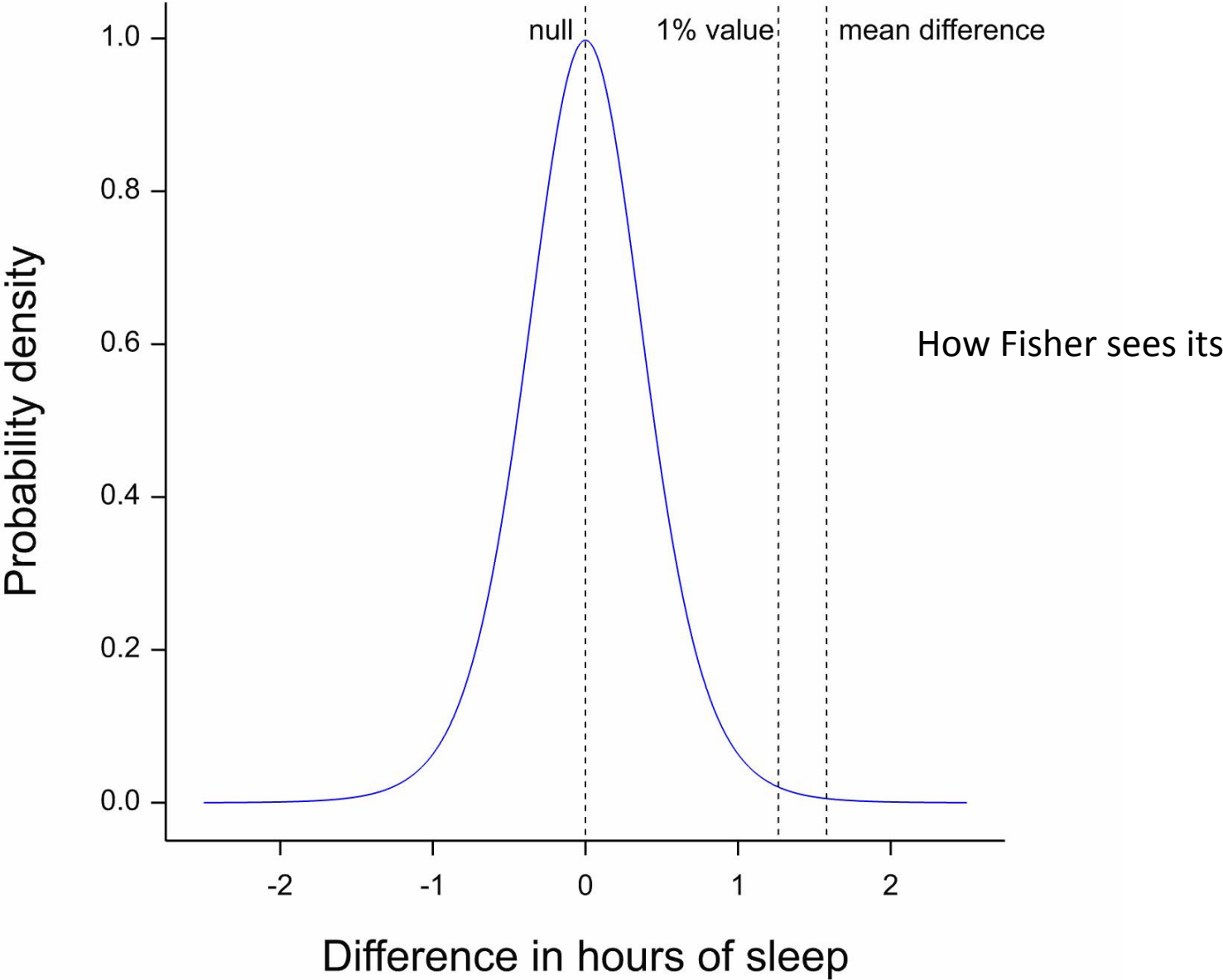
For  $n = 9$ , only one value in a hundred will exceed 3.250 by chance, so that the difference between the results is clearly significant.

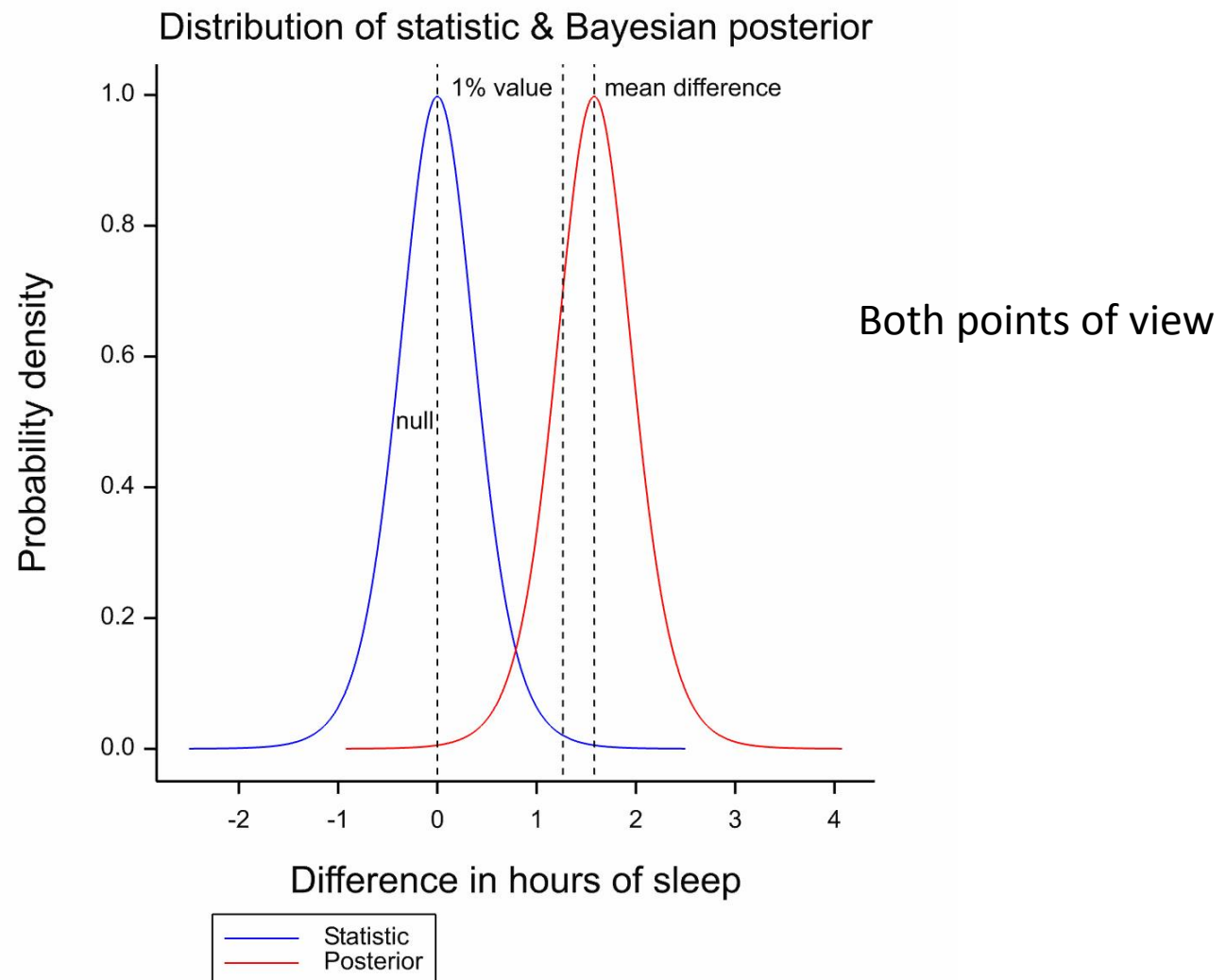
Fisher, *Statistical Methods for Research Workers*, 1925

# Bayesian posterior distribution



# Distribution of statistic under the null





# 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

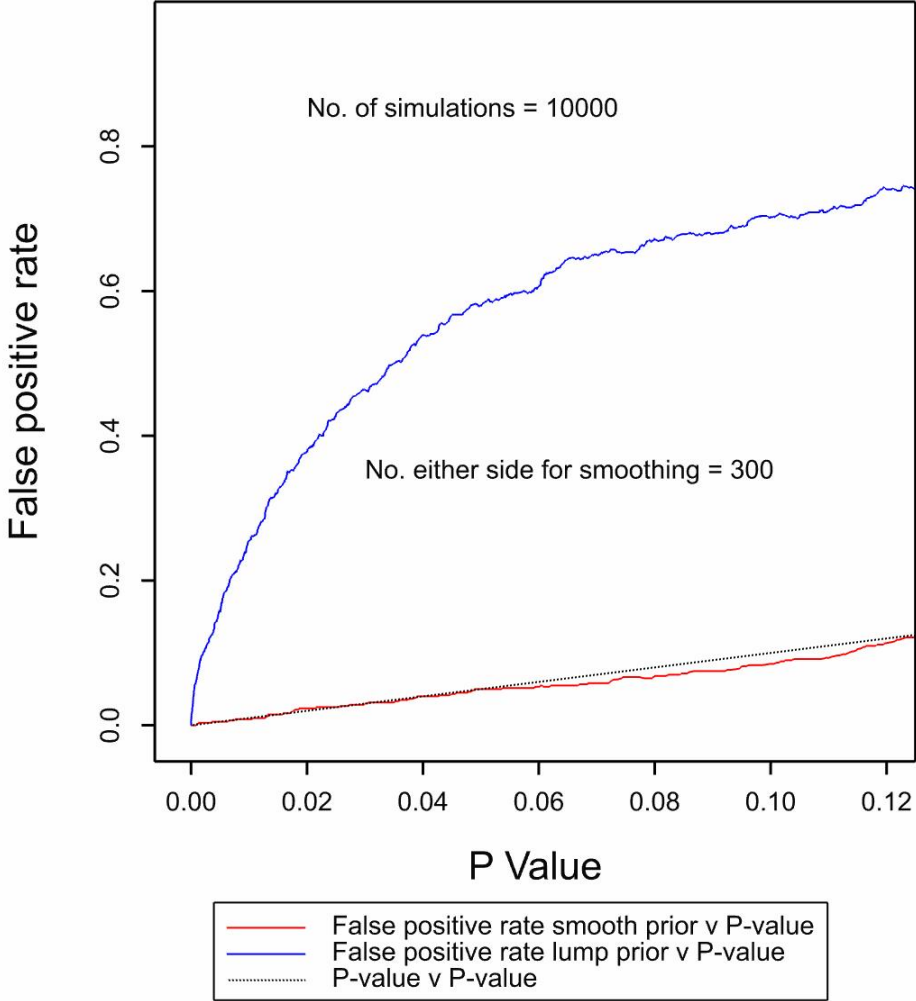
On these assumptions it can be proved that the probability that the *next* to be drawn will be white is  $\frac{m+1}{m+2}$ , and that the probability that *all* the  $n$  are white is  $\frac{m+1}{n+1}$ .

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



# Why the difference?

- Imagine a point estimate of two standard errors
- Now consider the likelihood ratio for a given value of the parameter,  $\delta$  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