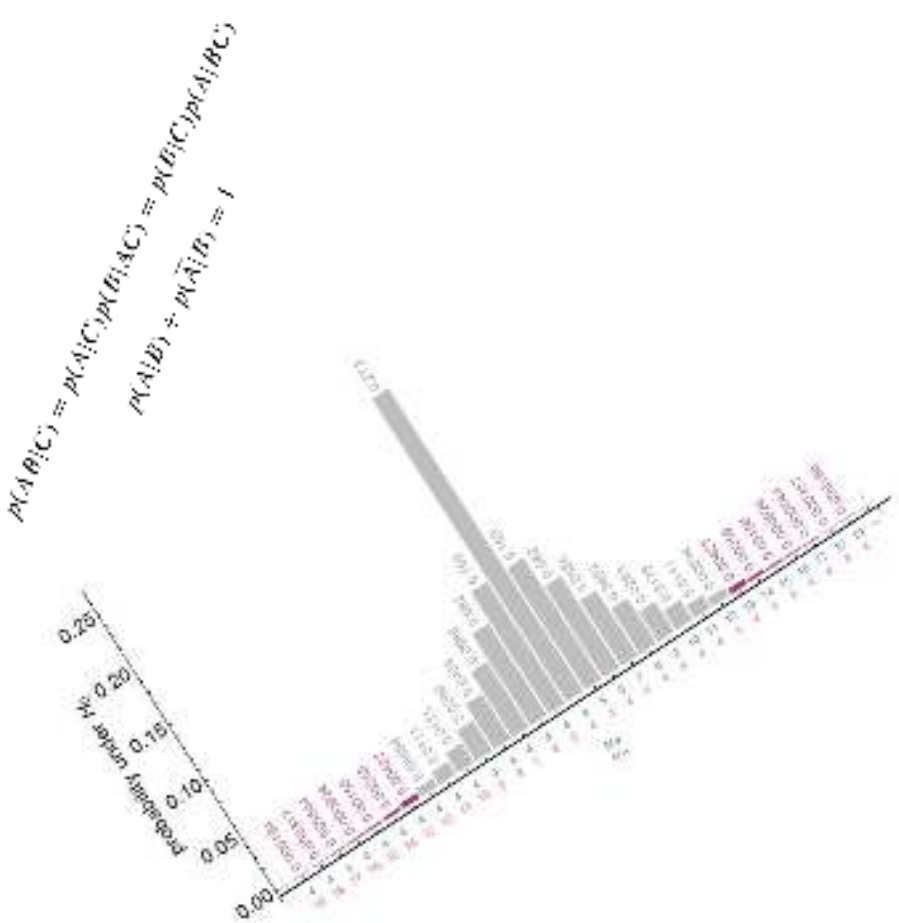
P(robability | introduction)

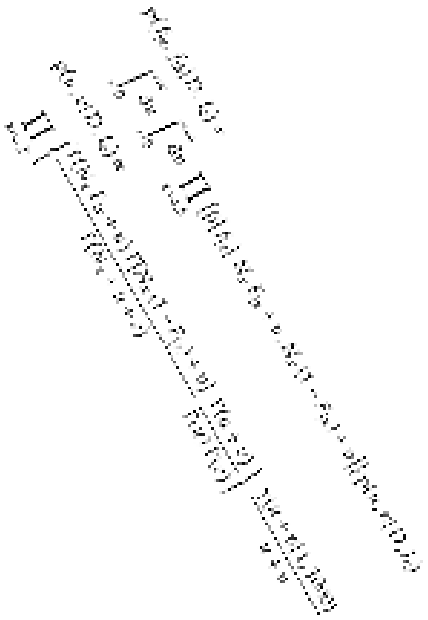
by Luca <piero.mana@ntnu.no>

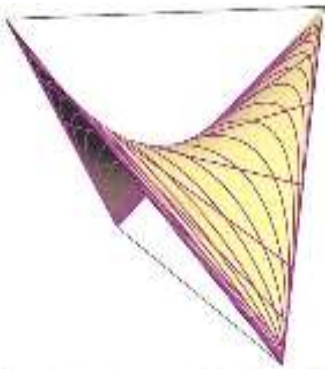
15 March 2119 (time machine)

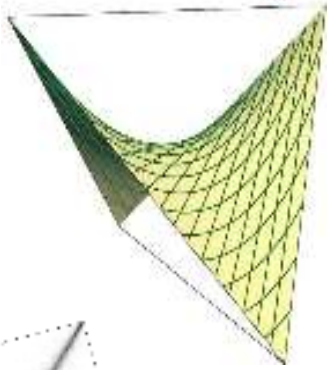
- Slides with text will be available in the course folder
- All cited literature is available in the course folder

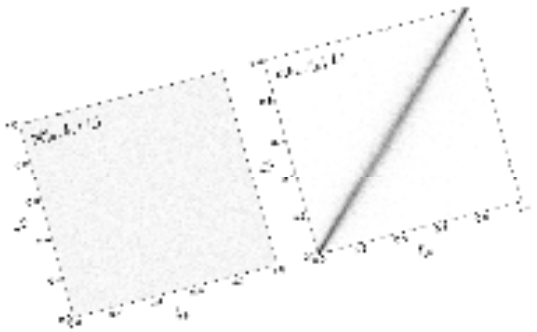
Probability theory











Probability theory uses many different mathematical concepts and techniques.

The purpose of these two lectures is to:
• explain the basic ideas behind *p*-values and null-hypothesis testing, and their problems;
• explain the basics of Bayesian probability theory.

We aim to understand how these work, not to become proficient with them.

People attending this course find themselves in a period of great change. There are two main ways to *think about* and *do* probability theory: they're usually called "frequentist" and "Bayesian".

Most scientific disciplines have moved or are moving away from the frequentist theory towards Bayesian theory. Astrophysics was among the first ones and today it only uses Bayesian theory. In genetics and neuroscience the shift has begun. It's therefore important to get acquainted with the new way, in order not to be left behind.

The transition happened because Bayesian theory gives demonstrably better results and because it's a *method*, based on a couple of principles only. Frequentist theory is a collection of recipes, of corrections to the recipes, and of dogmas pronounced by "authorities". No method there.

The main problem is that the two ways of doing probability do not differ only in formulae, but in the very *way of thinking* about probability. In some scientific problems they lead to the same conclusions; in some, to different conclusions; and in some, their results aren't comparable because they phrase and face the problem in completely different ways.

Probability theory

“Frequentist theory”:

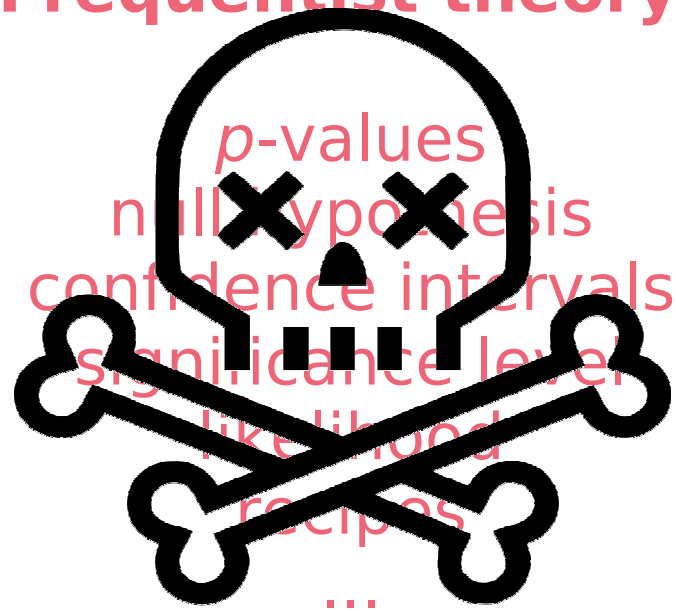
p-values
null hypothesis
confidence intervals
significance level
likelihood
recipes
...

“Bayesian theory”:

Bayes’s theorem
prior
posterior
loss function
maximum-entropy
method
...

Probability theory

“Frequentist theory”:



→

“Bayesian theory”:

Bayes’s theorem
prior
posterior
loss function
maximum-entropy
method
...

“Frequentist theory”:

probability = long-run[1] frequency

collection of recipes

“Bayesian theory”:

probability = degree of belief

method

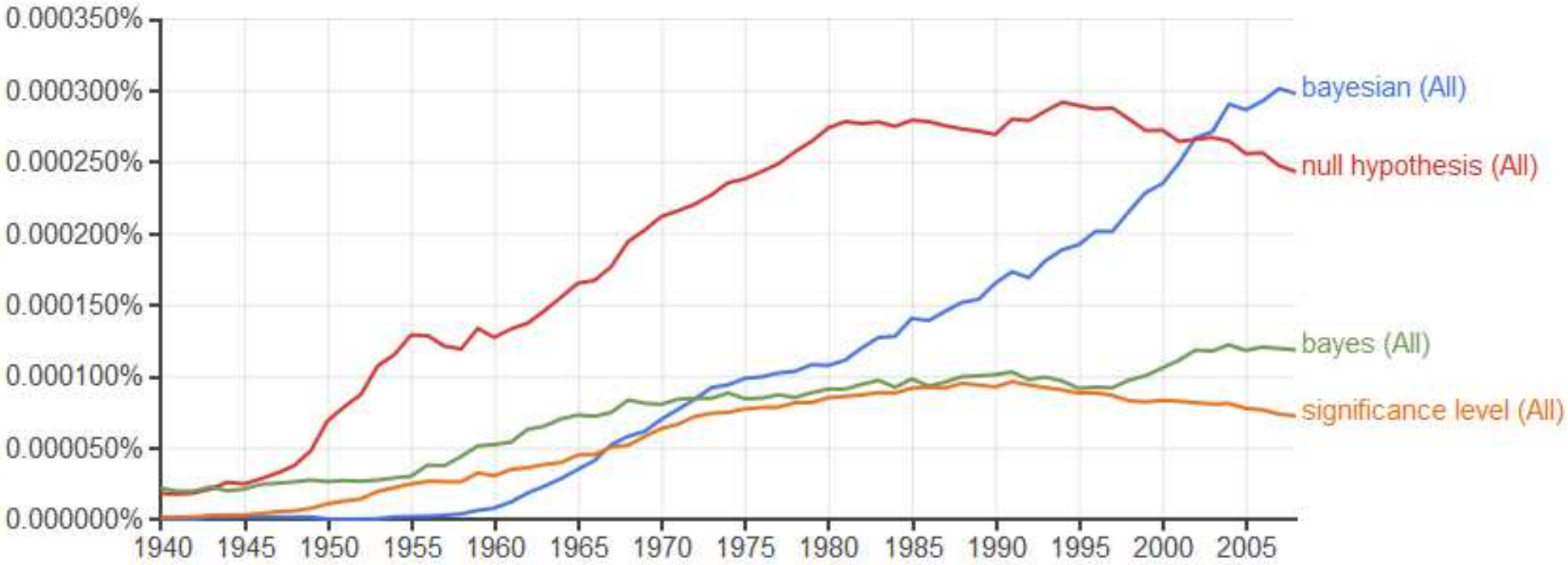
- different ways of thinking
- different results!

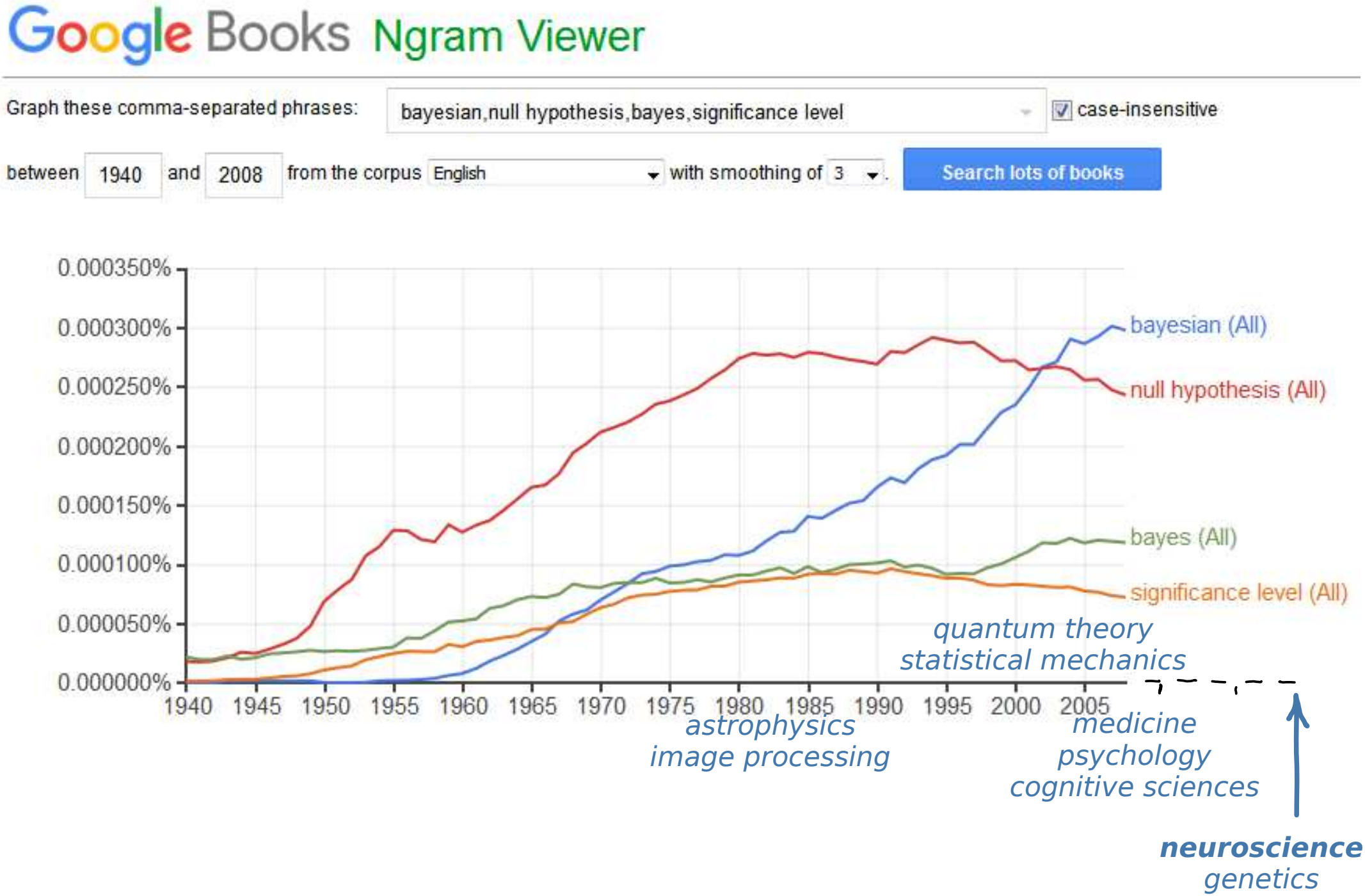
[1] “But this *long run* is a misleading guide to current affairs.
In the long run we are all dead” (J. M. Keynes 1923)

Google Books Ngram Viewer

Graph these comma-separated phrases: bayesian,null hypothesis,bayes,significance level ☑ case-insensitive

between 1940 and 2008 from the corpus English with smoothing of 3 . Search lots of books




Google Books Ngram Viewer
Graph these comma-separated phrases:
bayesian,null hypothesis,bayes,significance level
case-insensitive
between 1940 and 2008 from the corpus English with smoothing of 3 .
Search lots of books
0.000350%
0.000300%
0.000250%
0.000200%
0.000150%
0.000100%
0.000050%
0.000000%
1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005
bayesian (All)
null hypothesis (All)
bayes (All)
significance level (All)
quantum theory
statistical mechanics
astrophysics
image processing
medicine
psychology
cognitive sciences
neuroscience
genetics

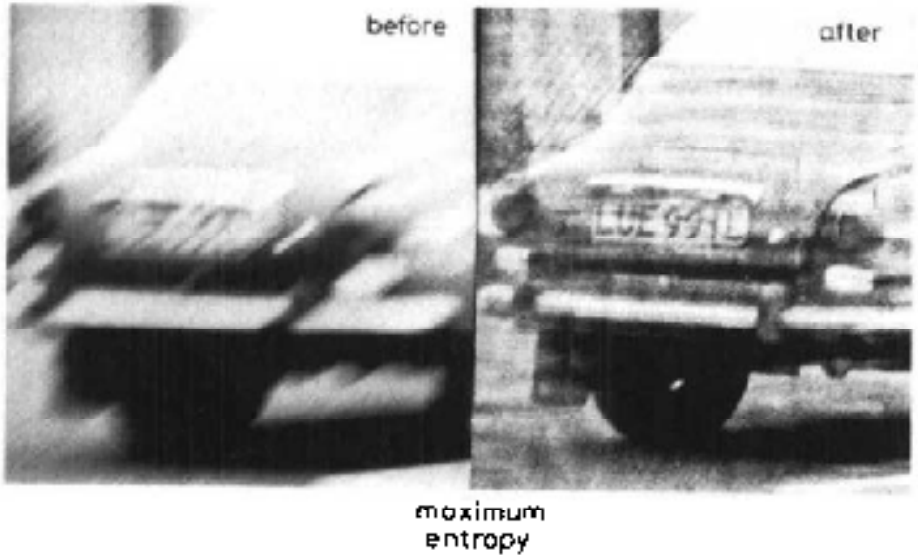


Fig. 1 *Photograph subject to motion blur of the camera*
a Point-spread-function is principally motion blur, with some out-of-focus component (UK Home Office photograph)
b Maximum entropy deconvolution

Maximum entropy method in image processing

S.F. Gull and J. Skilling

without discernible increase in resolution. This is as it should be, because the extra information within each resolution disc, which might have led to increased resolution, was simply lost in the noise. Maximum entropy gives an image with minimum structure, so that the balance between increased resolution and noise suppression is automatic. Fine structural details will appear in the reconstruction if and only if the data demand them.

Maximum entropy will also deal with incomplete data.

BAYESIAN SPECTRUM ANALYSIS ON QUADRATURE NMR DATA WITH NOISE CORRELATIONS (1989)

G. LARRY BRETTHORST

We then show that in typical NMR data the frequencies and decay rates may be estimated with a precision several orders of magnitude better than directly from the discrete Fourier transform.

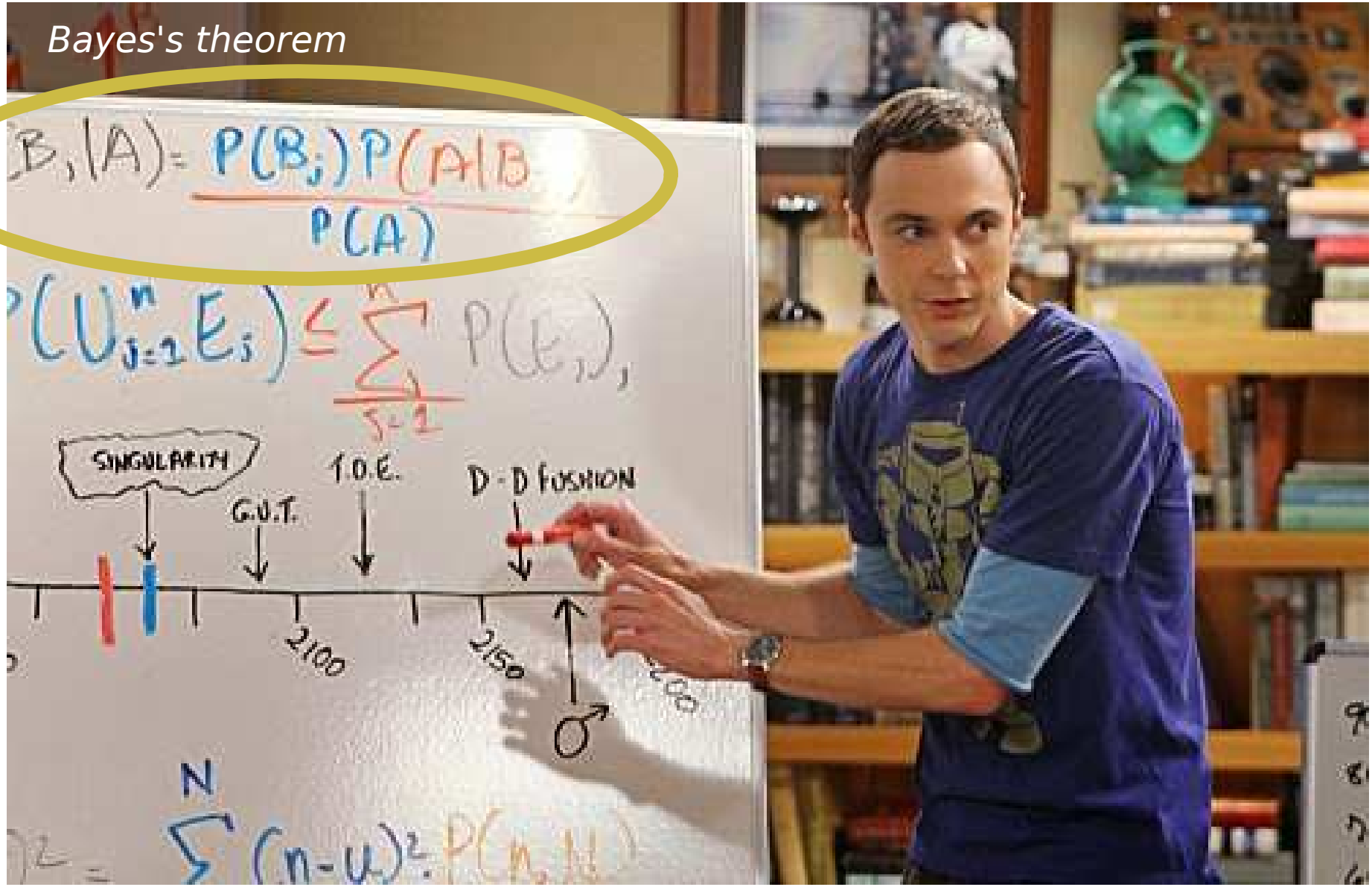
SATURDAY, NOVEMBER 10, 2012

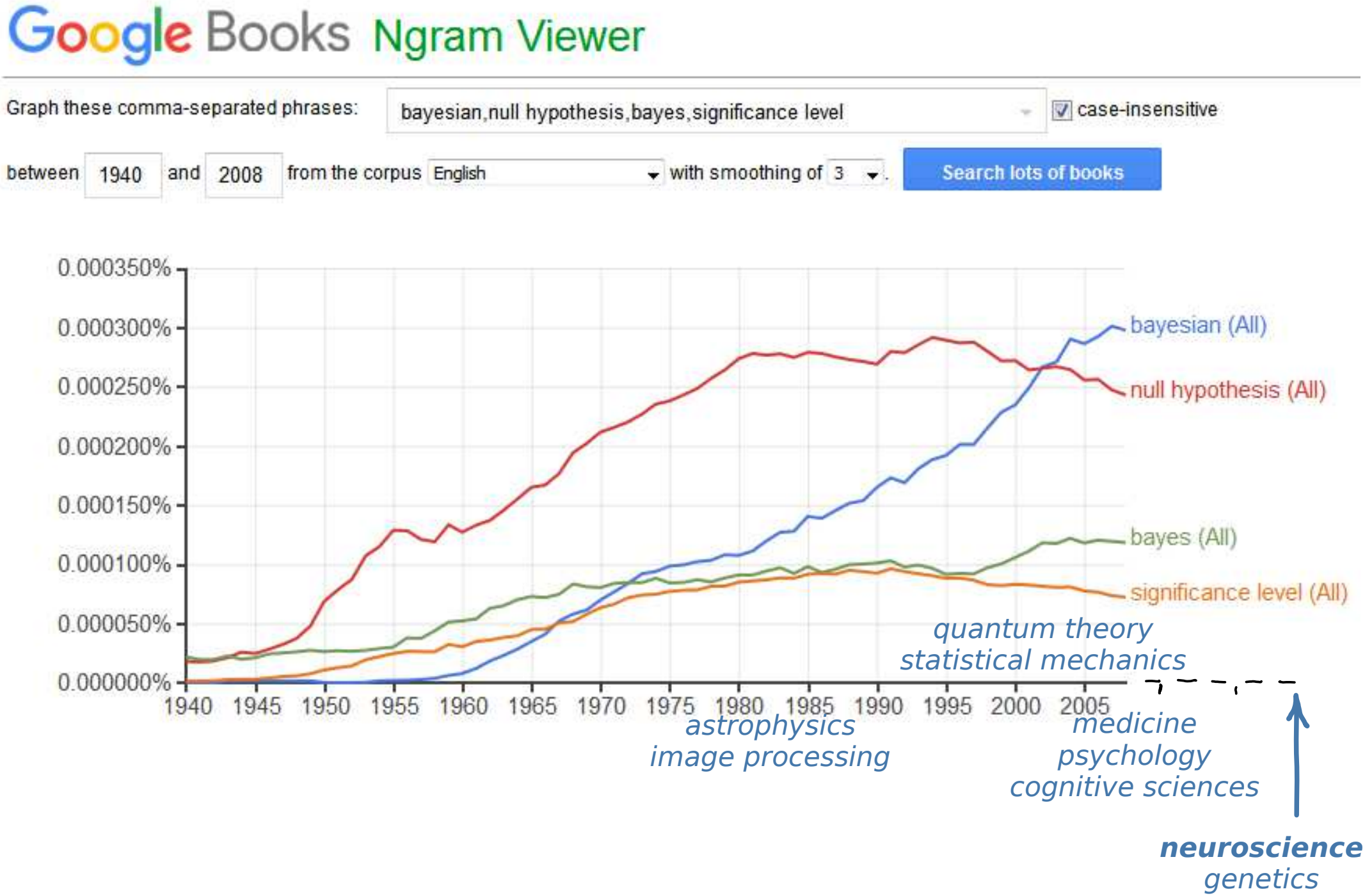
Nate Silver and the new Numerati

By now, we probably all know who Nate Silver is. He correctly forecast the result in 49 out of 50 states and all 35 US Senate Races in the 2008 election cycle and all 50 states in the 2012 election cycle. How did he do this? Bayesian Analysis. Ignore all the political pundits.. Nate simply removed the noise from the true signals. You can check out his 538 blog at the New York Times for more details.

Nate Silver's Map

The Actual Map


Bayes's theorem
P(B_i|A) = P(B_i)P(A|B_i) / P(A)
SINGULARITY
G.U.T.
T.O.E.
D-D FUSION
2100
2150


Google Books Ngram Viewer
Graph these comma-separated phrases:
bayesian,null hypothesis,bayes,significance level
case-insensitive
between 1940 and 2008 from the corpus English with smoothing of 3 .
Search lots of books
0.000350%
0.000300%
0.000250%
0.000200%
0.000150%
0.000100%
0.000050%
0.000000%
1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005
bayesian (All)
null hypothesis (All)
bayes (All)
significance level (All)
quantum theory
statistical mechanics
astrophysics
image processing
medicine
psychology
cognitive sciences
neuroscience
genetics

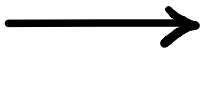
Frequentist theory:

p-values & null-hypothesis testing

To understand the idea behind null-hypothesis testing and *p*-values let's imagine to face the following investigation:

We're interested in the effects that a particular drug has on cognition in rats. It might lead to an increase or to a decrease of cognitive abilities, or leave cognition unaffected.

In null-hypothesis testing we usually choose *one* hypothesis (typically the "no-effect" one) and do experiments to guess if that hypothesis is false.



drug

(example adapted from Berger & al 1988)

cognitive +

drug

?

(example adapted from Berger & al 1988)

cognitive +

drug

?

cognitive -

(example adapted from Berger & al 1988)

cognitive +

drug

?

cognitive -

cognitive =

(example adapted from Berger & al 1988)

?

Null hypothesis H_0

Imagine this scenario: you visit a lab and find out that they're actually testing the drug. They've devised this experiment:

1. Two homozygote rat twins ard bred and raised in identical conditions.

2. At some point in their development one twin is chosen by the toss of a coin, and the drug injected into it. A placebo is injected in the other twin.

3. The twins undergo a cognitive task, devised in such a way that there's a "winner" and a "loser".

4. If the drugged twin is the winner, this is considered a "+" result. If it's the loser, this is considered a "–" result.

5. The procedure above is repeated for several twin pairs.

The idea is this: if the drug affects cognition, we should observe a prevalence of "+" over "–" or vice versa. If the drug has no effect, their numbers should be roughly equal, since the drugged twin is chosen by the toss of a coin.

[This imaginary experiment is likely to be poorly designed; feel free to imagine a better design with two outcomes. The point is to illustrate how *p*-values work.]

twins

one twin → drug

cognitive test:
which twin wins?

Repeat this test on several pairs of twins

You ask the experimenter in the lab what the outcome of the experiment was. They tell you that they tested 17 pairs of twins, observing 13 "+" and 4 "–".

Outcome: 17 pairs, 13 drug → cognition+, 4 drug → cognition–

P-value

1. List all possible outcomes that *could have been* obtained

P-value

1. List all possible outcomes that *could have been* obtained

2. Calculate the probability of every possible outcome under H_0

P-value

1. List all possible outcomes that *could have been* obtained

2. Calculate the probability of every possible outcome under H_0

3. Read the probability of the *actual* outcome

P-value

1. List all possible outcomes that *could have been* obtained

2. Calculate the probability of every possible outcome under H_0

3. Read the probability of the *actual* outcome

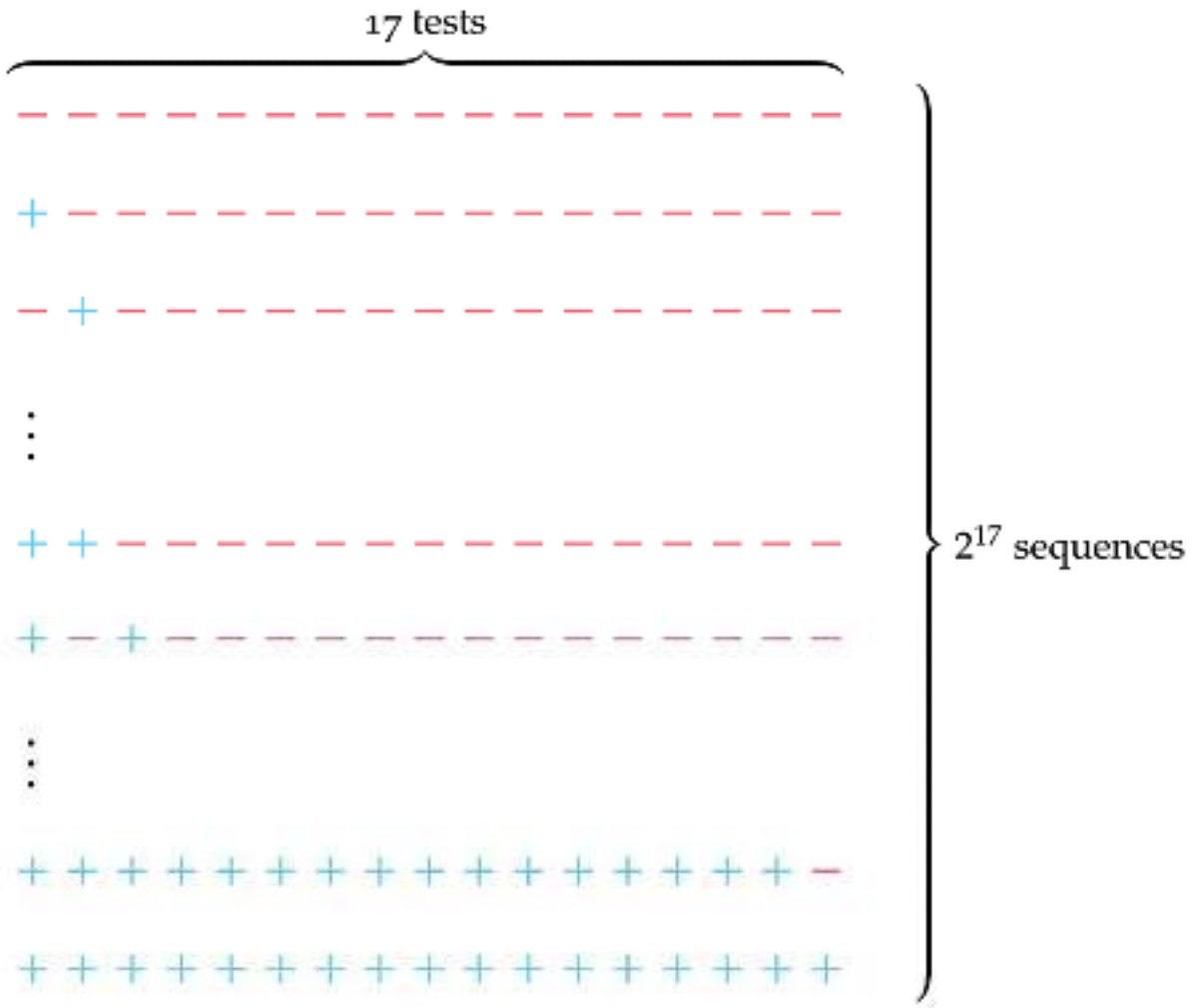
4. Sum the probabilities of all outcomes that have probability ≤ probability of actual outcome

→ That's the *p*-value

Probability of "possible observations that cast as much *or more* doubt on H_0 than do the actual data" (Berger & al 1988)

1. List all possible outcomes that *could have been* obtained

Possible *sequences* of test results:



Possible outcomes:

number of + : $N+$	=	0	1	2	3	4	5	⋯	12	13	14	15	16	17
number of − : $N-$		17	16	15	14	13	12		5	4	3	2	1	0

2. Calculate the probability of every possible outcome under H_0

Consider a specific outcome: $\begin{matrix} N+ \\ N- \end{matrix} = \begin{matrix} 13 \\ 4 \end{matrix}$

- Possible sequences for this outcome:

$- - - - + + + + + + + + + + + + + +$

$- - - + - + + + + + + + + + + + + +$

$- - + - - + + + + + + + + + + + + +$

$\vdots$

$- - + + - - + + + + + + + + + + + +$

$\vdots$

$+ + + + + + + + + + + + + + - - - -$

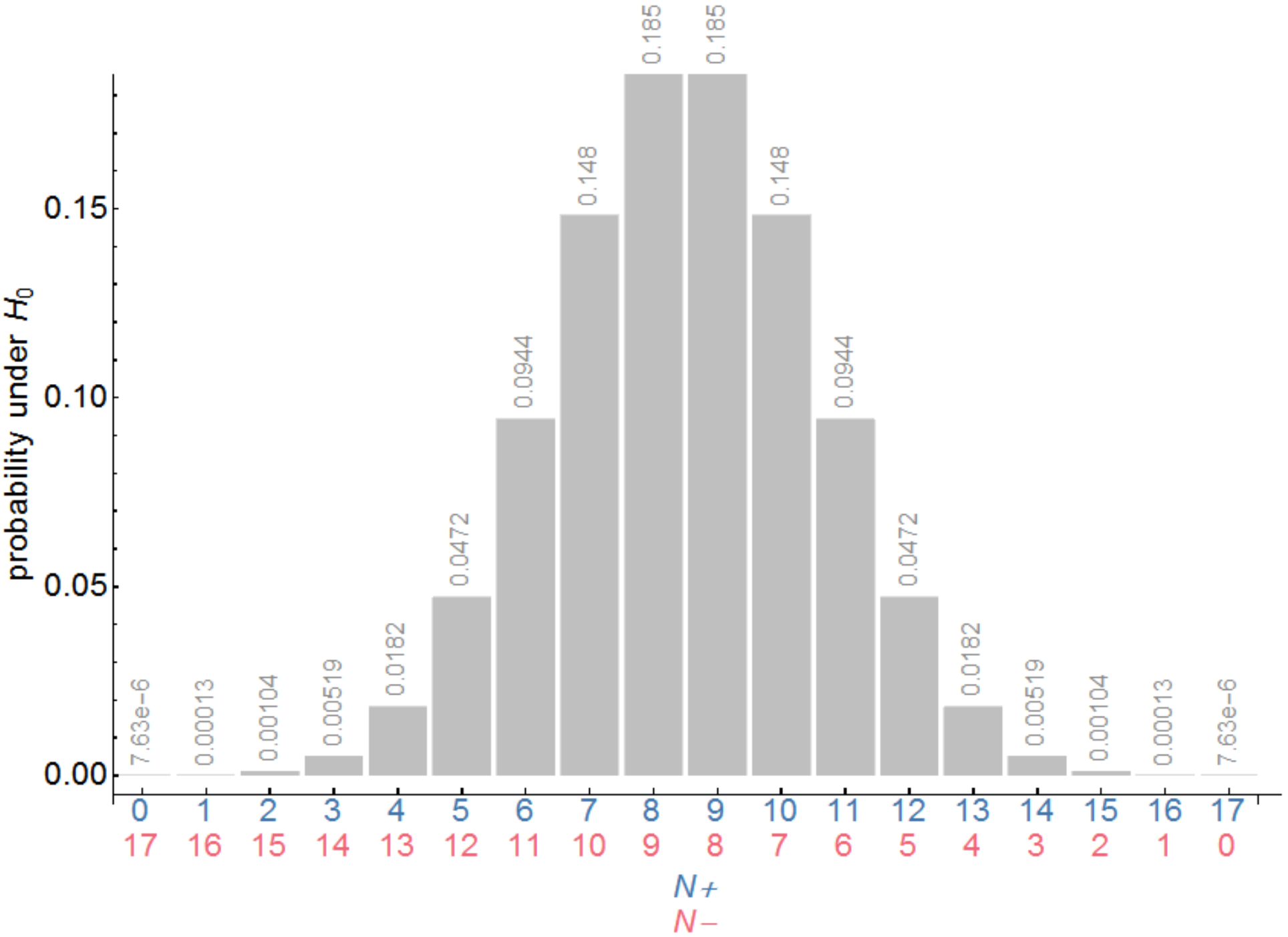
each one has probability $\frac{1}{2^{17}}$ under H_0

- How many? Number of distinct ways to shuffle 13 '+' and 4 '−':

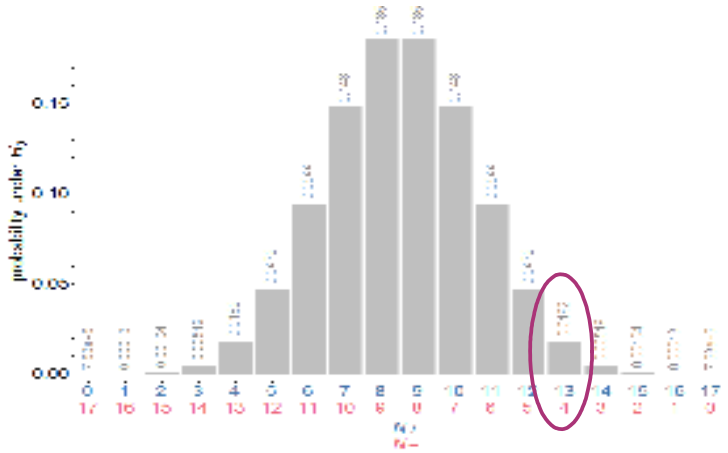
binomial coefficient

$$\binom{17}{13} = \binom{17}{4} = \frac{17!}{13!\,4!} = \frac{17 \times 16 \times \cdots \times 2 \times 1}{(13 \times 12 \times \cdots \times 2 \times 1)\,(4 \times 3 \times 2 \times 1)} = 2380$$

Outcome $\begin{matrix} 13 \\ 4 \end{matrix}$ **has probability** $\binom{17}{4} \times \frac{1}{2^{17}} = 0.0182$


probability under H_0
0.15
0.10
0.05
0.00
7.63e−6
0.00013
0.00104
0.00519
0.0182
0.0472
0.0944
0.148
0.185
0.185
0.148
0.0944
0.0472
0.0182
0.00519
0.00104
0.00013
7.63e−6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
N+
N−

3. Read the probability of the *actual* outcome

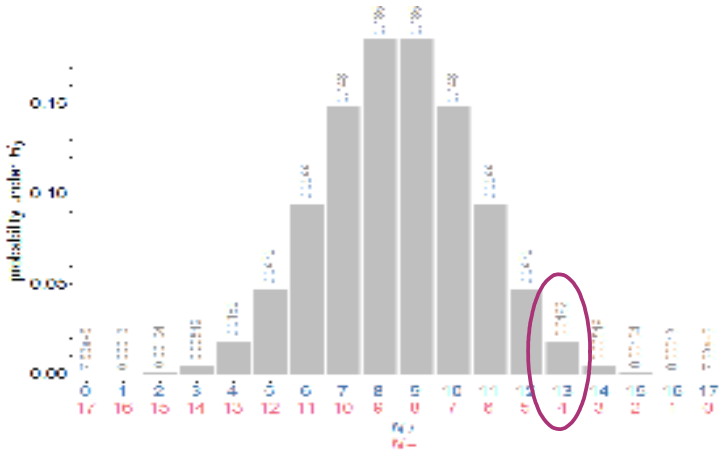


4. Sum the probabilities of all outcomes that have probability ≤ probability of actual outcome

→ That's the *p*-value

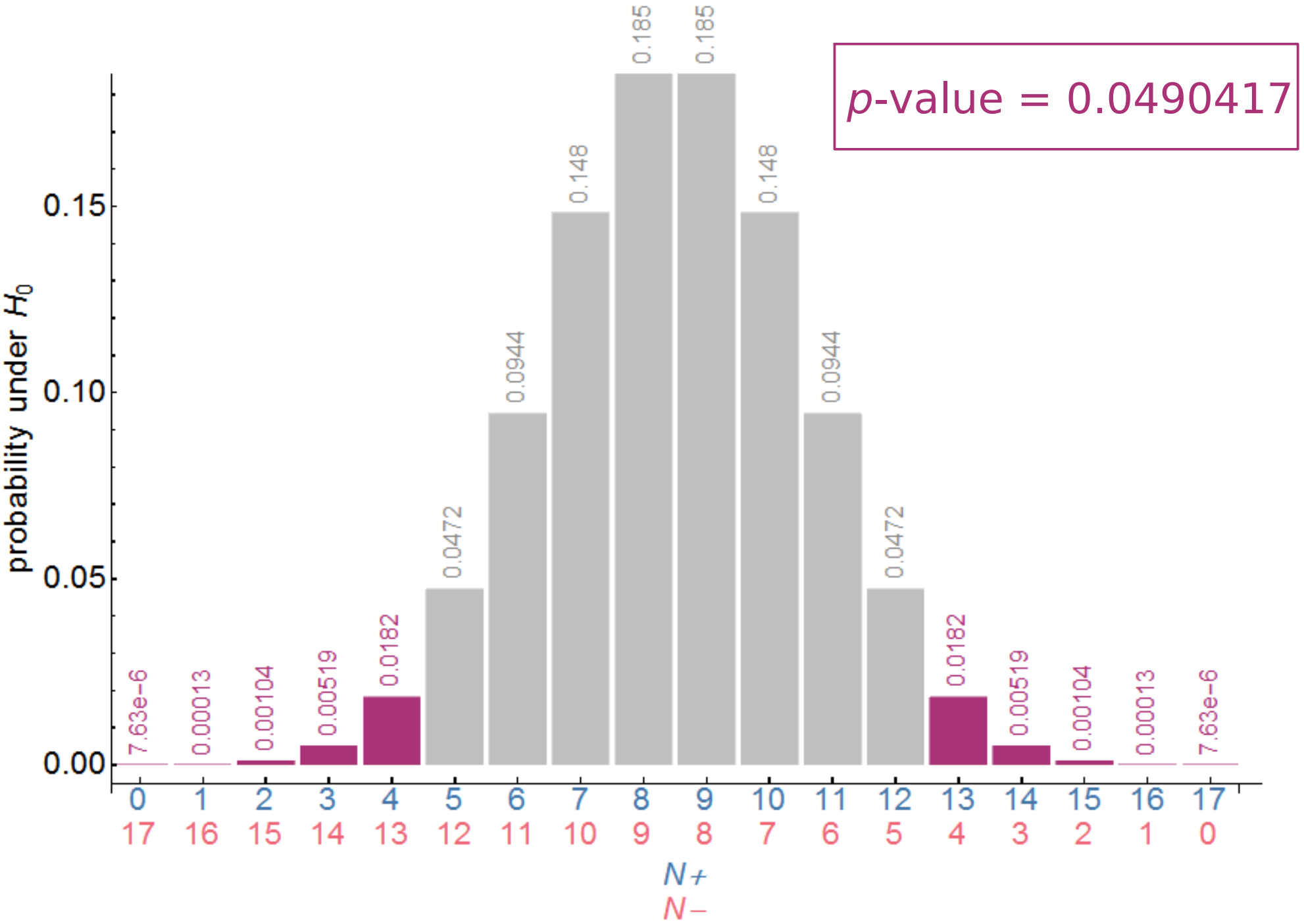
Competition! The first four will win...

3. Read the probability of the *actual* outcome



4. Sum the probabilities of all outcomes that have probability ≤ probability of actual outcome

→ That's the *p*-value


p-value = 0.0490417
probability under H_0
0.15
0.10
0.05
0.00
7.63e−6
0.00013
0.00104
0.00519
0.0182
0.0472
0.0944
0.148
0.185
0.185
0.148
0.0944
0.0472
0.0182
0.00519
0.00104
0.00013
7.63e−6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
N+
N−

Now imagine the following development in the scenario:

Some time later you visit another lab. It turns out that also in this lab they've been testing the drug – and that they used *the same exact* protocol as the first lab, up to every minute detail (the kind of rats, the environment, ...even the brand of syringes).

It turns out that the two labs *don't* know of each other's work (they haven't made their experiments and results public yet). Imagine that the researchers of the two labs don't even know one another.

...Surprise!

Essentially identical experiment was made in another lab

(same protocols, equivalent equipment & subjects, etc.)

This is a great coincidence, but nevertheless a coincidence. Intrigued by it, you ask lab#2 what their result was. They tell you that they tested 17 pairs of twins, observing 13 "+" and 4 "–" (it turns out that even the sequence of results is the same as lab#1's).

I must ask you to imagine that there hasn't been any exchange of information or interaction between the two labs. This is really just a coincidence.

lab#2: 17 pairs, 13 drug → cognition+, 4 drug → cognition–

Question:

Should the same p-value you found for lab #1 also apply to lab #2?

– Please ask me about any details that you judge important for answering this question –

YES	NO	DUNNO
24	2	3

...but most people changed their mind from YES to NO after the “why 17?” question

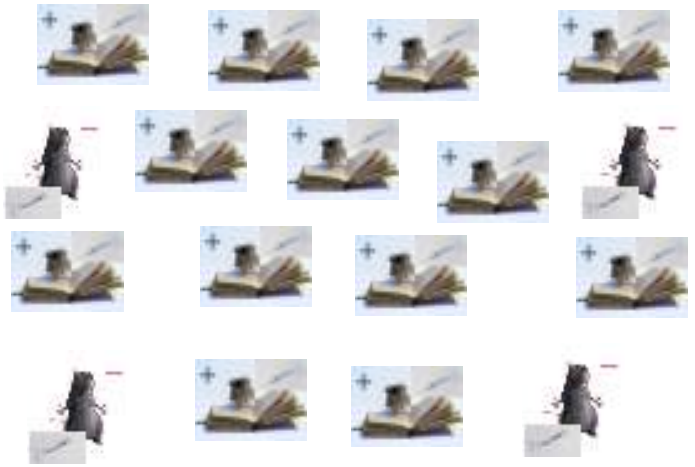
Your questions:

- Do the two labs know about each other's experiments? (No)
- Are we pooling the results of the two labs? (No)

Question: why was the number 17 chosen? (and not 16 or 18 etc)

- Experimenter lab #1:
  "Considering our resources (time, money…), I could test at most 17 pairs."

- Experimenter lab #2:
  "I wanted to make sure to have enough samples for the less frequent case ('+' or '–'), whatever it might be. So I decided to stop only when I had at least 4 samples for each case."



What about the *p*-value for Lab #2?

Lab #2

1. List all possible outcomes that *could have been* obtained

– Each case ('+' or '−') must appear at least 4 times
– At most one case may appear 4 times

Impossible sequences:

− + − − + + + + − + − + − − + − + − + − not enough tests

− + − − + + + + − + − + + − + − − − + − − + too many tests

Possible sequences:

− + − − + − + + − + − − + + + + + − + + − − − − − + − − +

Possible outcomes:

last test was +

$N+$ = … 4 … 4 4 4 … 4 4 5 … 12 13 14 … 100 …

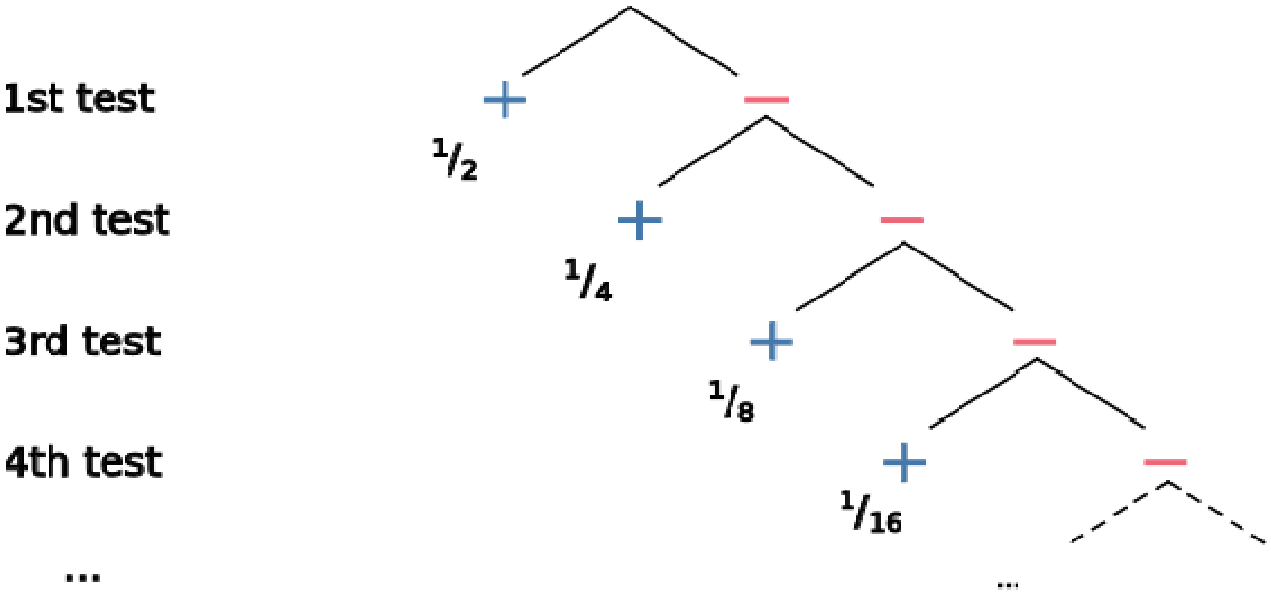
$N-$ = … 100 … 14 13 12 … 5 4 4 … 4 4 4 … 4 …

last test was −

Lab #2

2. Calculate the probability of every possible outcome under H_0

First: consider a simpler case: stop at first ‘+’



1/2 + 1/4 + 1/8 + 1/16 + ... = 1

Possible sequences	Probability
+	1/2
−+	1/4
−−+	1/8
−−−+	1/16
...	...
	= **1**

The probability
for each possible *sequence*
is the same
as **without** stopping rule
(even if some sequences
are now excluded)

Consider a specific case: $\begin{matrix} N+ \\ N- \end{matrix} = \begin{matrix} 13 \\ 4 \end{matrix}$

- Impossible sequences:

$$- - - - + + + + + + + + + + + + +$$

$$+ + - + + + + - - + + - + + + + +$$

- Possible sequences:

$$- - - + + + + + + + + + + + + + -$$

$$- - + + - + + + + + + + + + + + -$$

$$\vdots$$

$$+ + + + + + + + + + + + + + - - - -$$

each one has probability $\frac{1}{2^{17}}$ under H_0

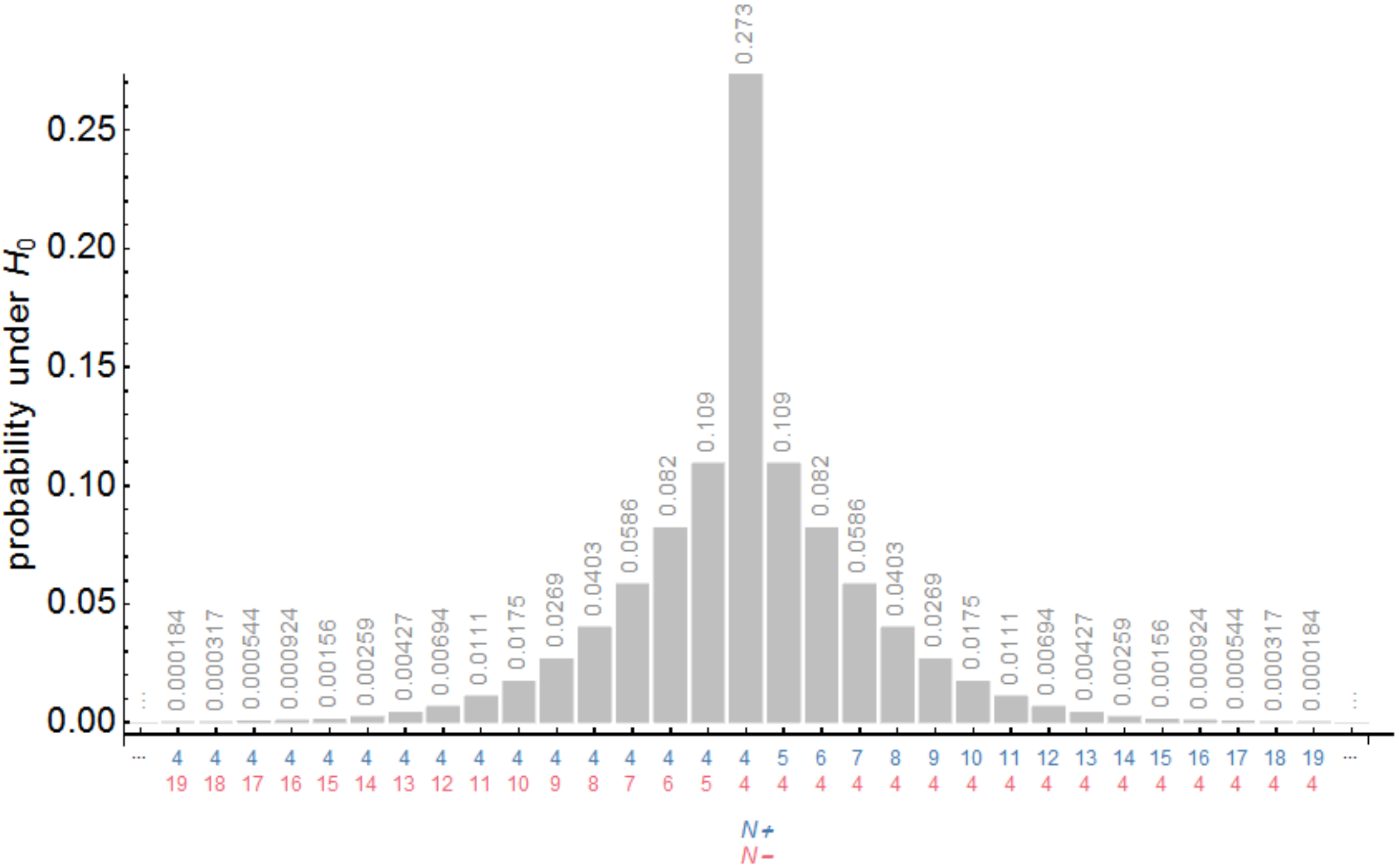
- How many? Number of distinct ways to shuffle 13 '+' and 3 '−':

binomial coefficient

$$\binom{16}{13} = \binom{16}{3} = \frac{16!}{13!\,3!} = \frac{16 \times 15 \times \cdots \times 2 \times 1}{(13 \times 12 \times \cdots \times 2 \times 1)\,(3 \times 2 \times 1)} = 560$$

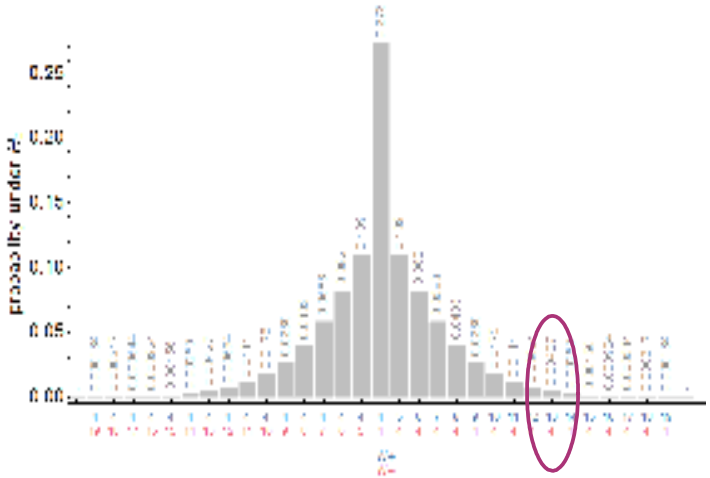
Outcome $\begin{matrix} 13 \\ 4 \end{matrix}$ **has probability** $\binom{16}{3} \times \frac{1}{2^{17}} = 0.00427$

Lab #2


probability under H_0
0.25
0.20
0.15
0.10
0.05
0.00
0.000184
0.000317
0.000544
0.000924
0.00156
0.00259
0.00427
0.00694
0.0111
0.0175
0.0269
0.0403
0.0586
0.082
0.109
0.273
0.109
0.082
0.0586
0.0403
0.0269
0.0175
0.0111
0.00694
0.00427
0.00259
0.00156
0.000924
0.000544
0.000317
0.000184
... 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 ...
... 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 ...
N+
N−

Lab #2

3. Read the probability of the *actual* outcome

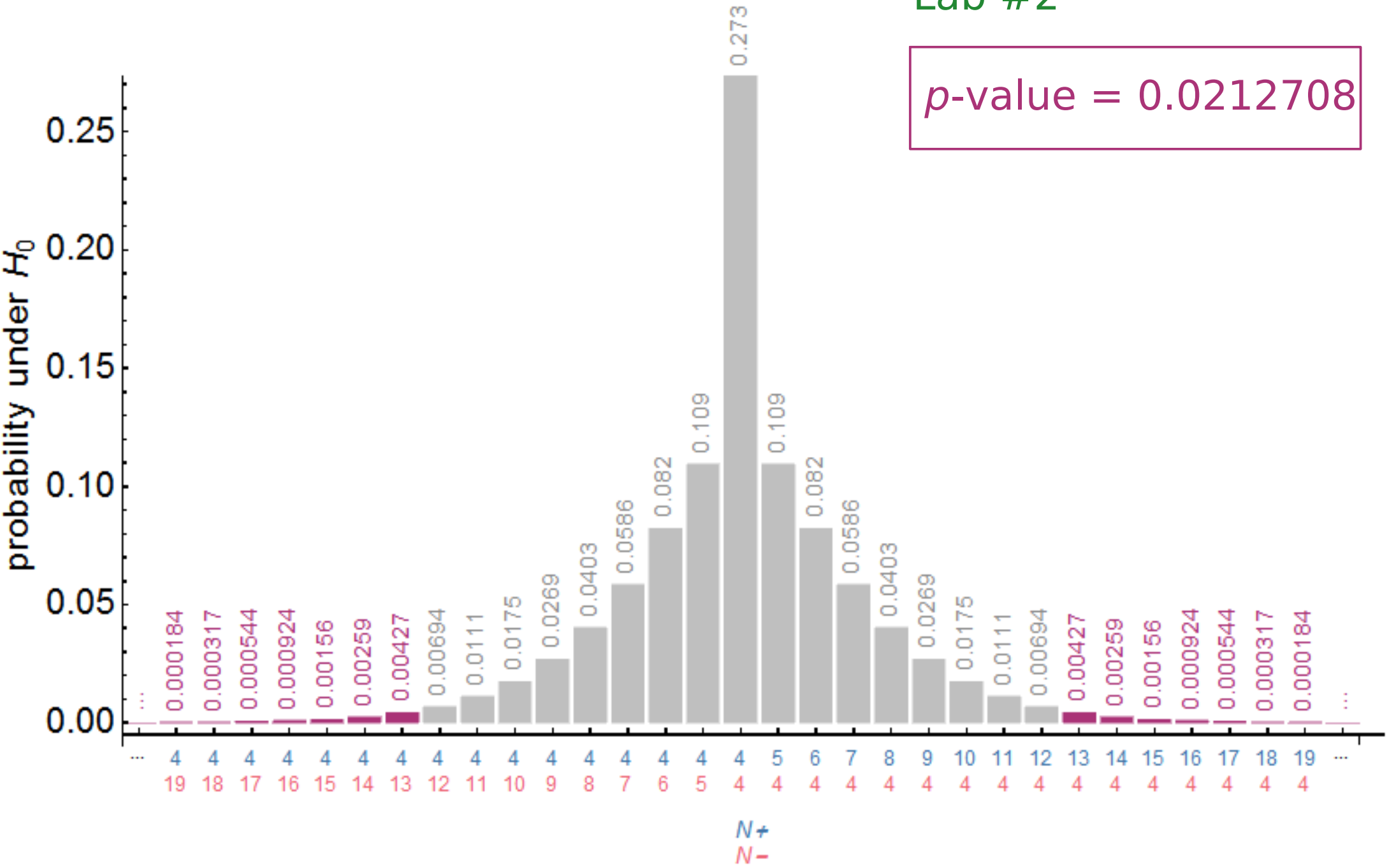


4. Sum the probabilities of all outcomes that have probability ≤ probability of actual outcome

→ That's the *p*-value

Lab #2

p-value = 0.0212708



The surprise is that even if the two labs used completely identical protocols and materials, and even if they got exactly the same outcome (even in the same order), the calculation of *p*-value leads to two different results.

Lab #1

outcome: 13 + 4 –

p-value for H_0: **0.049**

Lab #2

outcome: 13 + 4 –

p-value for H_0: **0.021**

Lab #1

outcome: 13 + 4 –

p-value for H_0: **0.049**

Lab #2

outcome: 13 + 4 –

p-value for H_0: **0.021**

Using other stopping rules,
the difference in p-values can be made as large as we please

(Anscombe 1954, Berger & al 1988, Wagenmakers 2006, ...)

LAB #1

This was the 17th pair.
Time to stop.

Now I have 4 of each.
Time to stop.

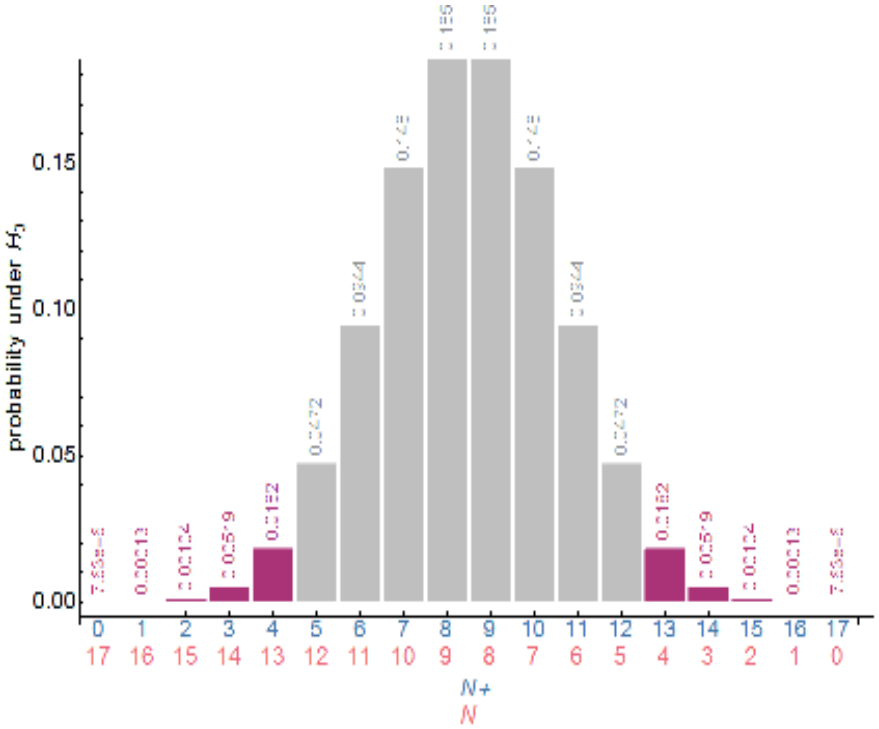
LAB #2

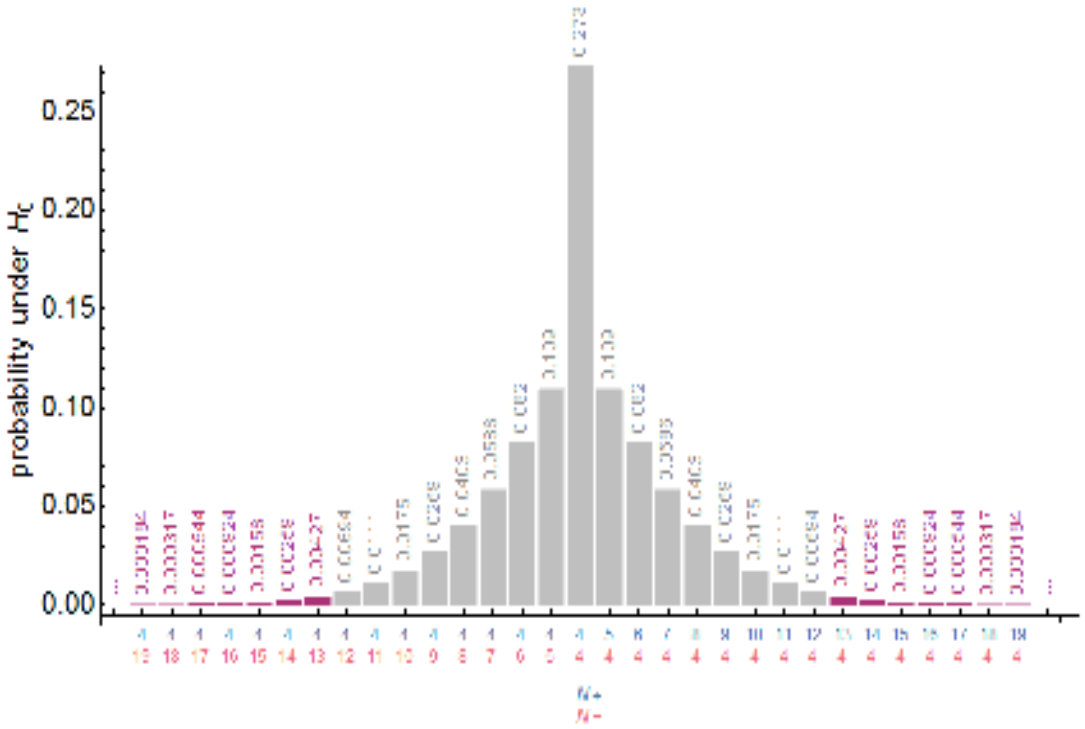
What happened in the two labs could have been *completely identical* except only for the *inner thoughts* of the two experimenters!

What happened in the two labs could have been *completely identical* except only for the *inner thoughts* of the two experimenters!

Telekinetic dependence – why?

Because the *p*-value depends on outcomes that *could have* occurred – but didn't





Probability for sequence that did *actually* occur = 0.00000763 for both labs

THEORY OF PROBABILITY

BY

HAROLD JEFFREYS

SECOND EDITION
OXFORD
AT THE CLARENDON PRESS
1948

(§ 7.2) *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it. The same applies to all the current significance tests based on P integrals.

Another example: a funny story from...

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley, C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

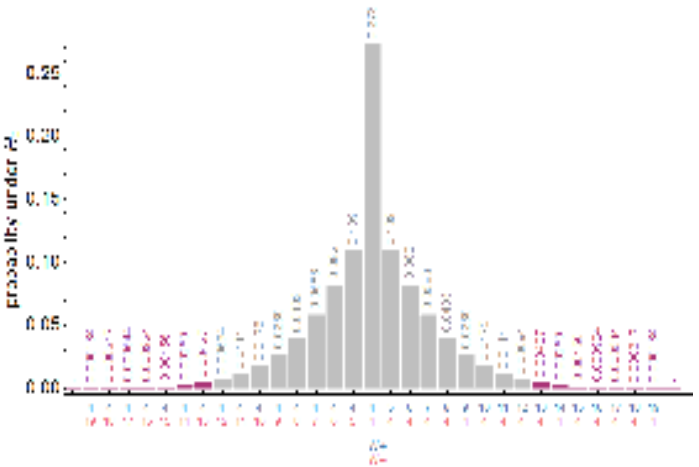
JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley,
C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate volt-meter, accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean.



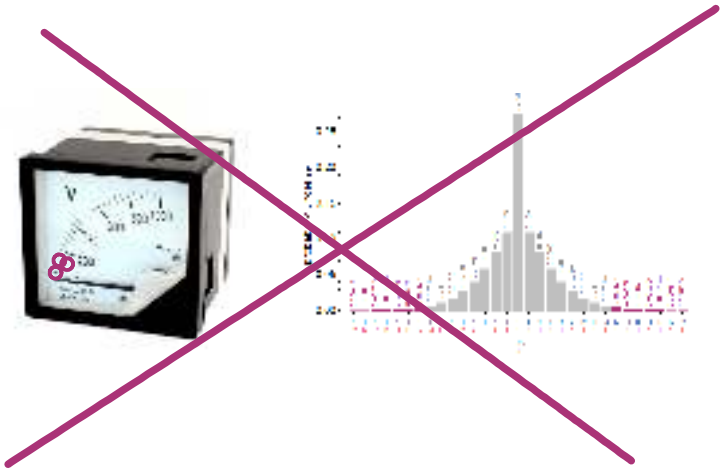
JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

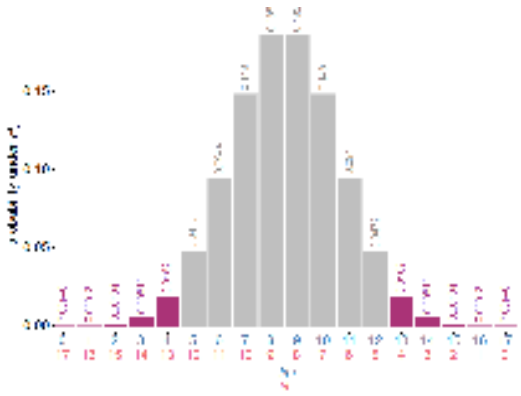
Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley,
C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

Later he visits the engineer's laboratory, and notices that the volt-meter used reads only as far as 100, so the population appears to be "censored." This necessitates a new analysis, if the statistician is orthodox.





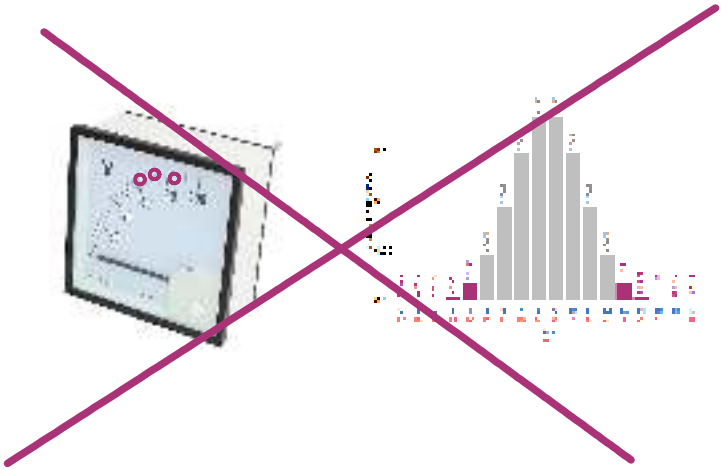
JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

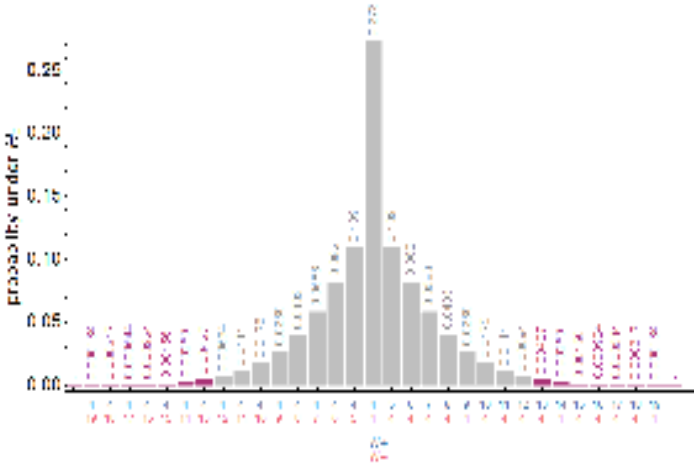
Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley,
C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

However, the engineer says he has another meter, equally accurate and reading to 1000 volts, which he would have used if any voltage had been over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all.





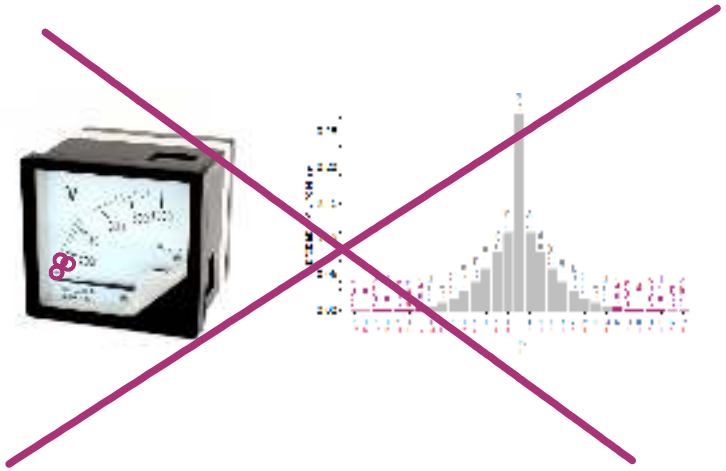
JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

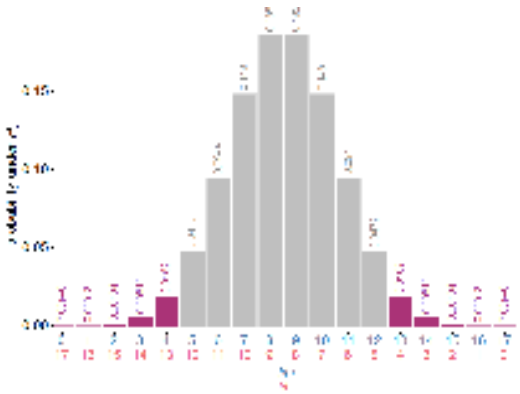
Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley,
C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

But the next day the engineer telephones and says, "I just discovered my high-range volt-meter was not working the day I did the experiment you analyzed for me." The statistician ascertains that the engineer would not have held up the experiment until the meter was fixed, and informs him that a new analysis will be required.





JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley,
C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

The engineer is astounded. He says, "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope."

JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley,
C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

The engineer is astounded. He says, "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope."

JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE

A. Birnbaum, L. J. Savage, G. Barnard, J. Cornfield, I. Bross, G. E. P. Box, I. J. Good, D. V. Lindley,
C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne

The engineer is astounded. He says, "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope."

I agree with the engineer. If the sample has voltages under 100, it doesn't matter whether the upper limit of the meter is 100, 1000, or 1 million. The sample provides the same information in any case.

P-values and other frequentist methods suffer from this dependence on details that our intuition tells us shouldn't be relevant.

This is not a rare situation, but a concrete occurrence in research. For example, imagine that you've applied for a grant for this particular experiment, and you should be informed soon on whether you won the grant.

You decide to test 10 subjects first, and if in the meantime you're notified that you won the grant, then you'll test 10 more.

If you now *exactly* apply the four steps to calculate the *p*-value for your experiment, you realize that you must consider the probability that you'd win the grant. It's strange that this probability should play a part in quantifying whether your null-hypothesis is true.

More generally, researchers often don't decide on the number of samples beforehand. This affects the *p*-value calculation. Most statistical software implicitly assumes that you decided the number of samples beforehand – so they're actually calculating the wrong *p*-value.

See the article by Wagenmakers (2017) for other very realistic examples of this quirky behaviour of *p*-values.

For other examples see:

Psychonomic Bulletin & Review
2007, 14 (5), 779-804

THEORETICAL AND REVIEW ARTICLES

A practical solution to the pervasive problems of *p* values

ERIC-JAN WAGENMAKERS
University of Amsterdam, Amsterdam, The Netherlands

In the field of psychology, the practice of *p* value null-hypothesis testing is as widespread as ever. Despite this popularity, or perhaps because of it, most psychologists are not aware of the statistical peculiarities of the *p* value procedure. In particular, *p* values are based on data that were never observed, and these hypothetical data are themselves influenced by subjective intentions. Moreover, *p* values do not quantify statistical evidence. This article reviews these *p* value problems and illustrates each problem with concrete examples. The three problems are familiar to statisticians but may be new to psychologists. A practical solution to these *p* value problems is to adopt a model selection perspective and use the Bayesian information criterion (BIC) for statistical inference (Raftery, 1995). The BIC provides an approximation to a Bayesian hypothesis test, does not require the specification of priors, and can be easily calculated from SPSS output.

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

The results were "plain as day", recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The *P* value, a common index for the strength of evidence, was 0.01 — usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the *P* value came out as 0.59 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame[1].

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false[2]; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results.

THE INSIGNIFICANCE OF STATISTICAL SIGNIFICANCE TESTING

DOUGLAS H. JOHNSON,[1] U.S. Geological Survey, Biological Resources Division, Northern Prairie Wildlife Research Center, Jamestown, ND 58401, USA

WHY ARE HYPOTHESIS TESTS USED?

With all the deficiencies of statistical hypothesis tests, it is reasonable to wonder why they remain so widely used. Nester (1996) suggested several reasons: (1) they appear to be objective and exact; (2) they are readily available and easily invoked in many commercial statistics packages; (3) everyone else seems to use them; (4) students, statisticians, and scientists are taught to use them; and (5) some journal editors and thesis supervisors demand them.

More cynically, Carver (1978) suggested that complicated mathematical procedures lend an air of scientific objectivity to conclusions. Shaver (1993) noted that social scientists equate being quantitative with being scientific. D. V. Lindley (quoted in Matthews 1997) observed that "People like conventional hypothesis tests because it's so easy to get significant results from them."

Other flaws:

- It’s possible to *reject* a *true* H_0 with as low p-level as desired
  (Anscombe 1954, Kadane & al 1996, ...)
- One hypothesis only (if it’s rejected, what’s left?)
- All possible hypotheses can end up being rejected
- Problem when used with rare events
  (some scientific disciplines mainly test rare events)
- Problem with small sample size
- **Easy to misinterpret & misuse**

Other flaws:

- It's possible to *reject* a *true* H_0 with as low p-level as desired
- One hypothesis only (if it's rejected, what's left?)
- All possible hypotheses can end up being rejected
- Problem when used with rare events
  (some scientific disciplines mainly test rare events)
- Small sample size
- Easy to misinterpret & misuse

Appl. Statist. (1996)
45, *No.* 4, *pp.* 401–410

An Applied Statistician's Creed

By MARKS R. NESTER†

Queensland Forestry Research Institute, Gympie, Australia

[Received March 1994. Final revision June 1996]

SUMMARY
Hypothesis testing, as performed in the applied sciences, is criticized. Then assumptions that the author believes should be axiomatic in all statistical analyses are listed. These assumptions render many hypothesis tests superfluous. The author argues that the image of statisticians will not improve until the nexus between hypothesis testing and statistics is broken.

nature International weekly journal of science

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Rights & Permissions

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

nature human behaviour

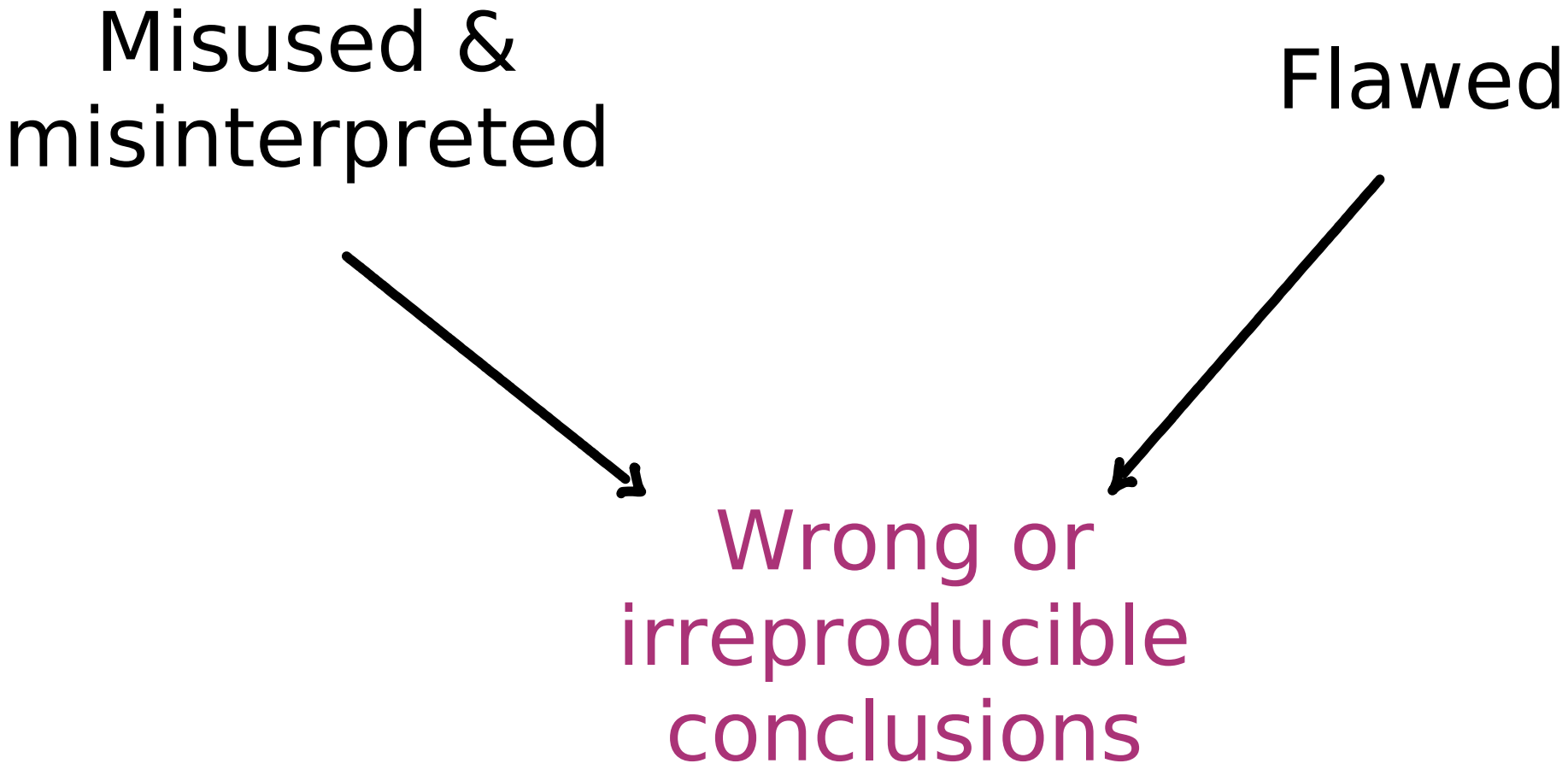
LETTERS

https://doi.org/10.1038/s41562-018-0399-z

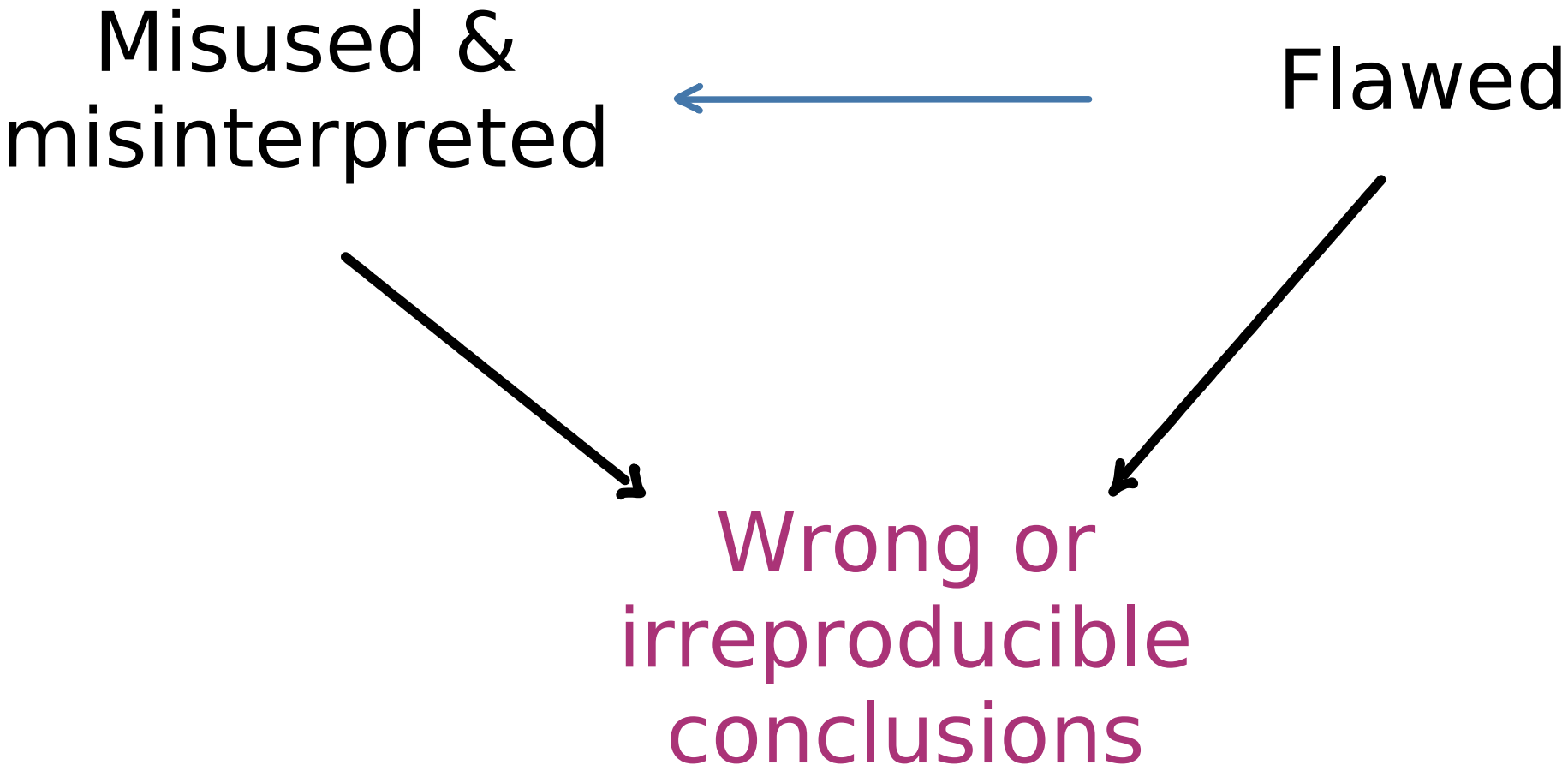
Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1]*, Anna Dreber[2,9], Felix Holzmeister[3,9], Teck-Hua Ho[4,9], Jürgen Huber[3], Magnus Johannesson[2], Michael Kirchler[5,9], Gideon Nave[6], Brian A. Nosek[7,8]*, Thomas Pfeiffer[9], Adam Altmejd[2], Nick Buttrick[7], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]

p-values & significance tests:



p-values & significance tests:



With the increasing use of statistical software, researchers don't consult statisticians any longer. And they end up miscalculating and misinterpreting *p*-values. This has lead to unreliable and contradictory conclusions in the literature.

In my opinion, researchers misuse the *p*-value also because the *p*-values is flawed, but they don't know that it is. Who'd imagine that their calculation might depend on whether they drank coffee in the morning? it sounds absurd and unscientific.

The statisticians have been very alarmed by the misuse of *p*-values, and even more by they over-interpretation as "significance".

For this reason the American Statistical Association published an official declaration in 2016, warning *against p*-values.

A year before, a mainstream psychology journal *banned* the use of *p*-values and null-hypothesis testing, because of their flaws.

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

American Statistical Association

ASA Statement on Statistical Significance and *P*-Values

EDITORIAL

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.**

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.**

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

※ Summary ※

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.**
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.**

- More reliable methods exist

“Homework”

- If you use statistical software, check whether it asks about the stopping rule of your experiment.
- Make a list of the stopping rules you’ve come across in your research (even vague ones).
- Calculate the p-value of concrete or simplified data. Imagine a different stopping rule. Re-calculate it and check how much it differs.