Why Bayes is Right and Everything Else is Wrong

EJ Wagenmakers
University of Amsterdam

Why Bayes is ~~Right~~ and Everything Else is ~~Wrong~~

EJ Wagenmakers
University of Amsterdam

Why Bayes Makes Sense and Everything Else is Silly

EJ Wagenmakers
University of Amsterdam

Why Bayes ~~Makes Sense~~ and Everything Else is ~~Silly~~

EJ Wagenmakers
University of Amsterdam

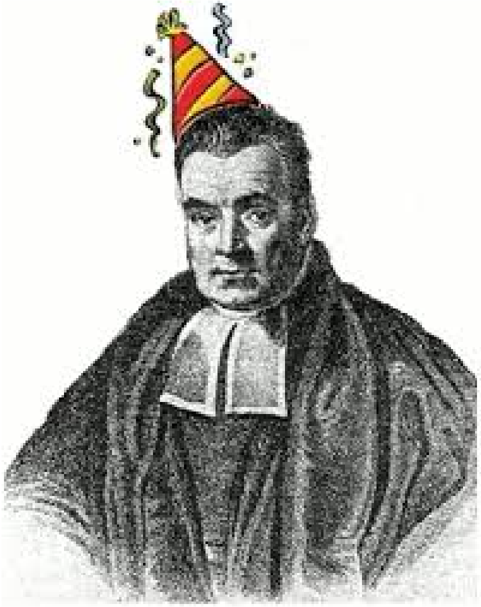
Why Bayes is Beautiful and Everything Else is Ugly

EJ Wagenmakers
University of Amsterdam

Why Bayes is ~~Beautiful~~ and Everything Else is ~~Ugly~~

EJ Wagenmakers
University of Amsterdam

Why Bayes Sparks Joy and Almost Everything Else Sparks Much Less Joy, in my Personal Experience

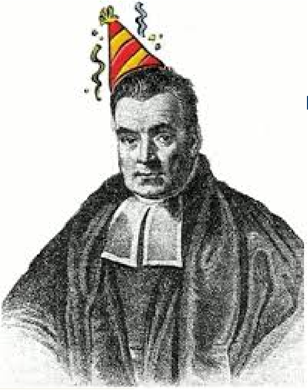


EJ Wagenmakers
University of Amsterdam

Bio

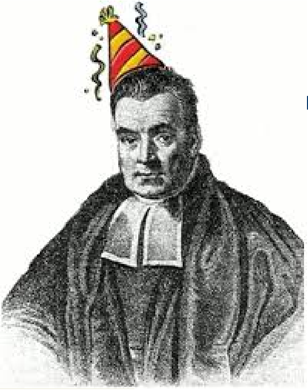
- Psychological Methods Unit @ UvA
- Main interests:
  - Bayesian inference
  - Open-source statistical software (JASP)
  - The *Journal of Robustness Reports*

JOURNAL OF
ROBUSTNESS
REPORTS



Outline

- What is Bayesian inference?
- Current popularity
- Unique advantages
- Errors: Type B and Type D
- Bayesian hypothesis testing
- Conclusion

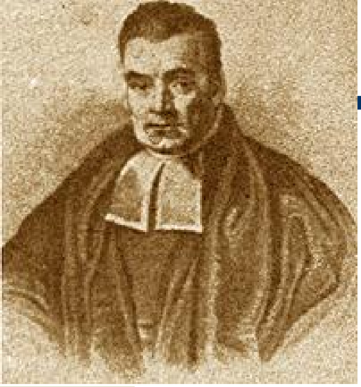


Outline

- What is Bayesian inference?
- Current popularity
- Unique advantages
- Errors: Type B and Type D
- Bayesian hypothesis testing
- Conclusion

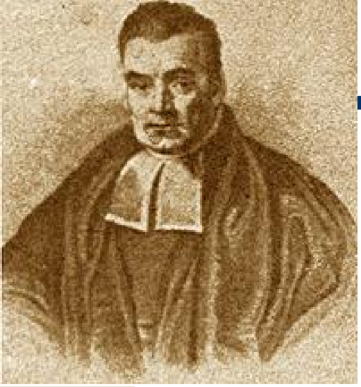
What is Bayesian Inference?

"Common sense expressed in numbers"



Bayesian Inference in a Nutshell

- In Bayesian inference, uncertainty or degree of belief is quantified by probability.
- Prior uncertainty is continually updated by means of the data to yield posterior uncertainty.



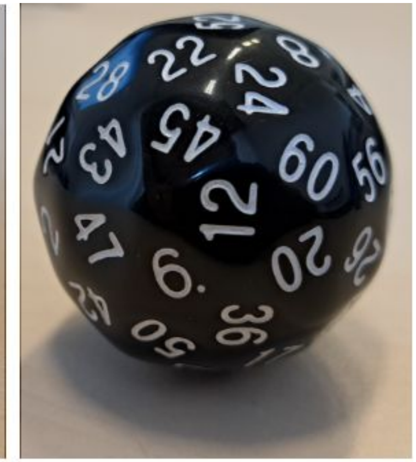
Bayesian Inference in a Nutshell

Hypotheses that predicted the data well enjoy a boost in credibility, whereas hypotheses that predicted the data poorly suffer a decline.

D3

D6

D12

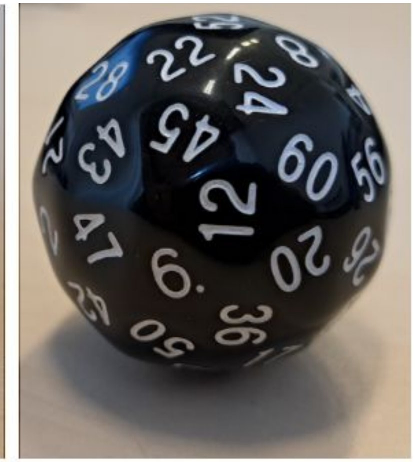


D60


D3


D6


D12


D60

You see the following outcomes:

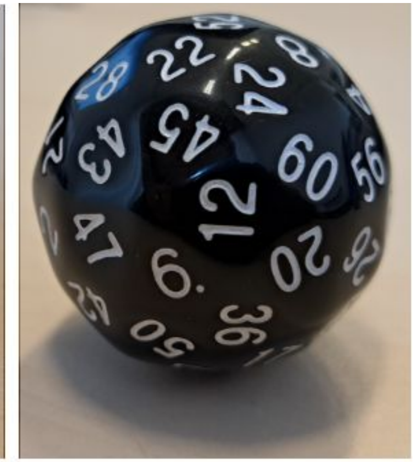
2, 3, 3

What die do you think generated these outcomes?

D3

D6

D12



D60

You see five more outcomes:

2, 3, 3, 3, 1, 1, 2, 3

What die do you think generated these outcomes? Are you more confident now?

D3 D6 D12 D60

You see five more outcomes:

2, 3, 3, 3, 1, 1, 2, 3

What die do you think generated these outcomes? Are you more confident now?

"Common sense expressed in numbers"

D3 2 D60

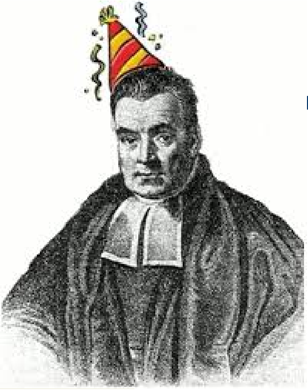
You s utcomes:

2, 3

What die c nerated these

outcomes? confident now?

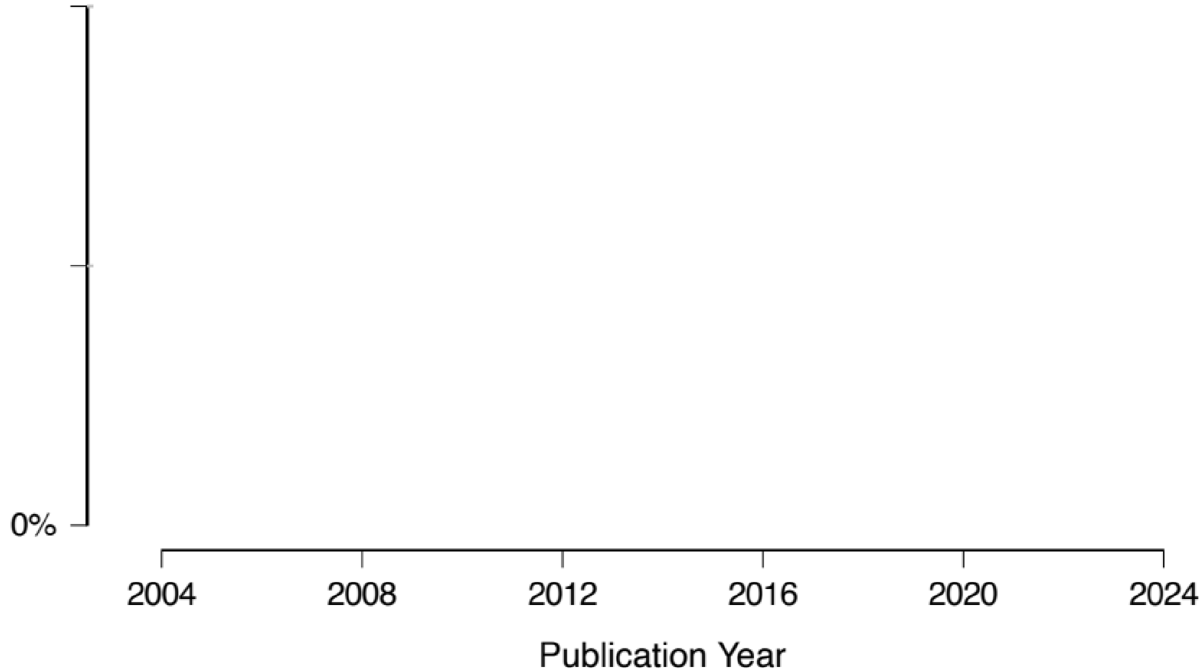
“[Common sense expressed in numbers]”

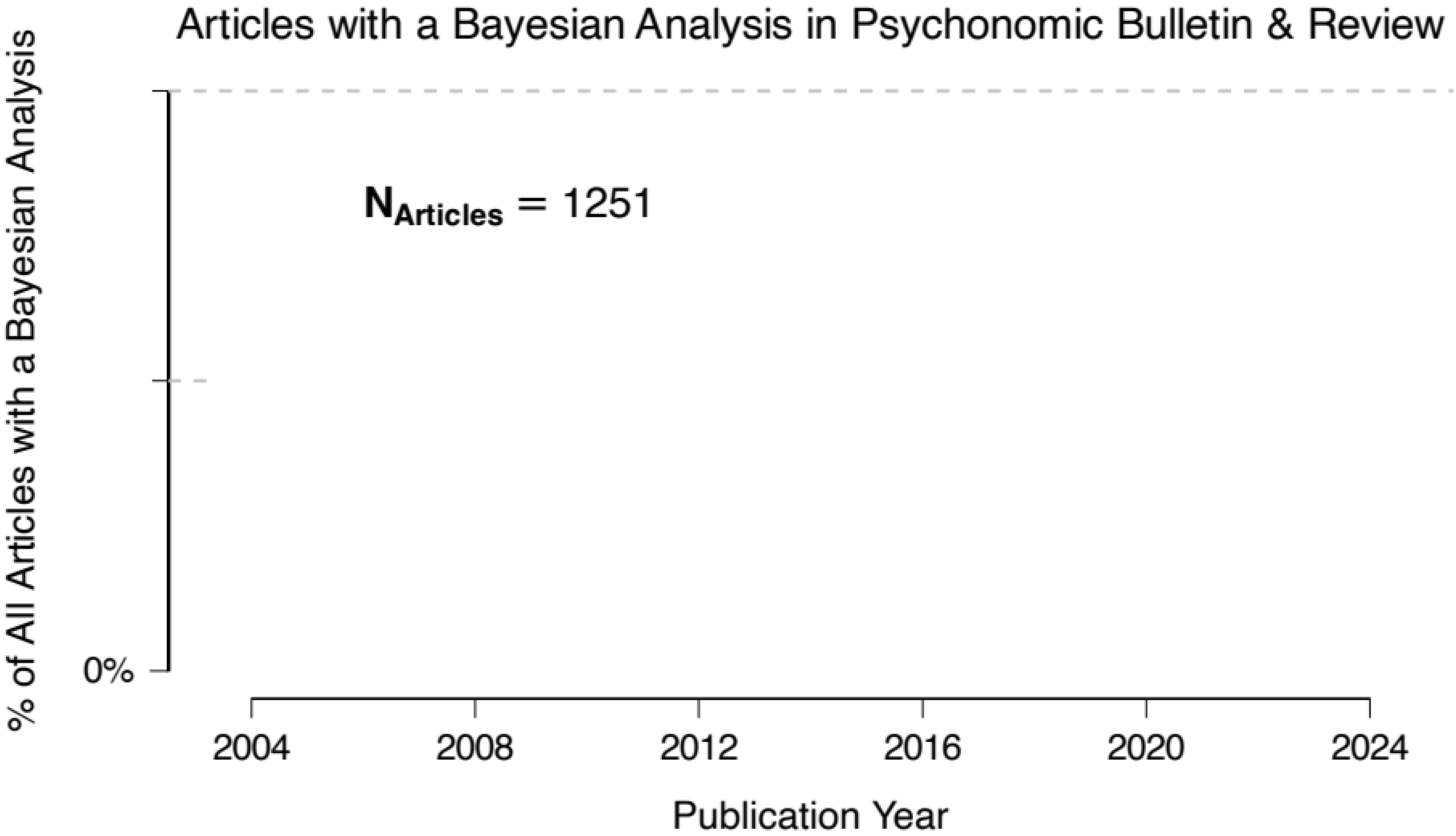


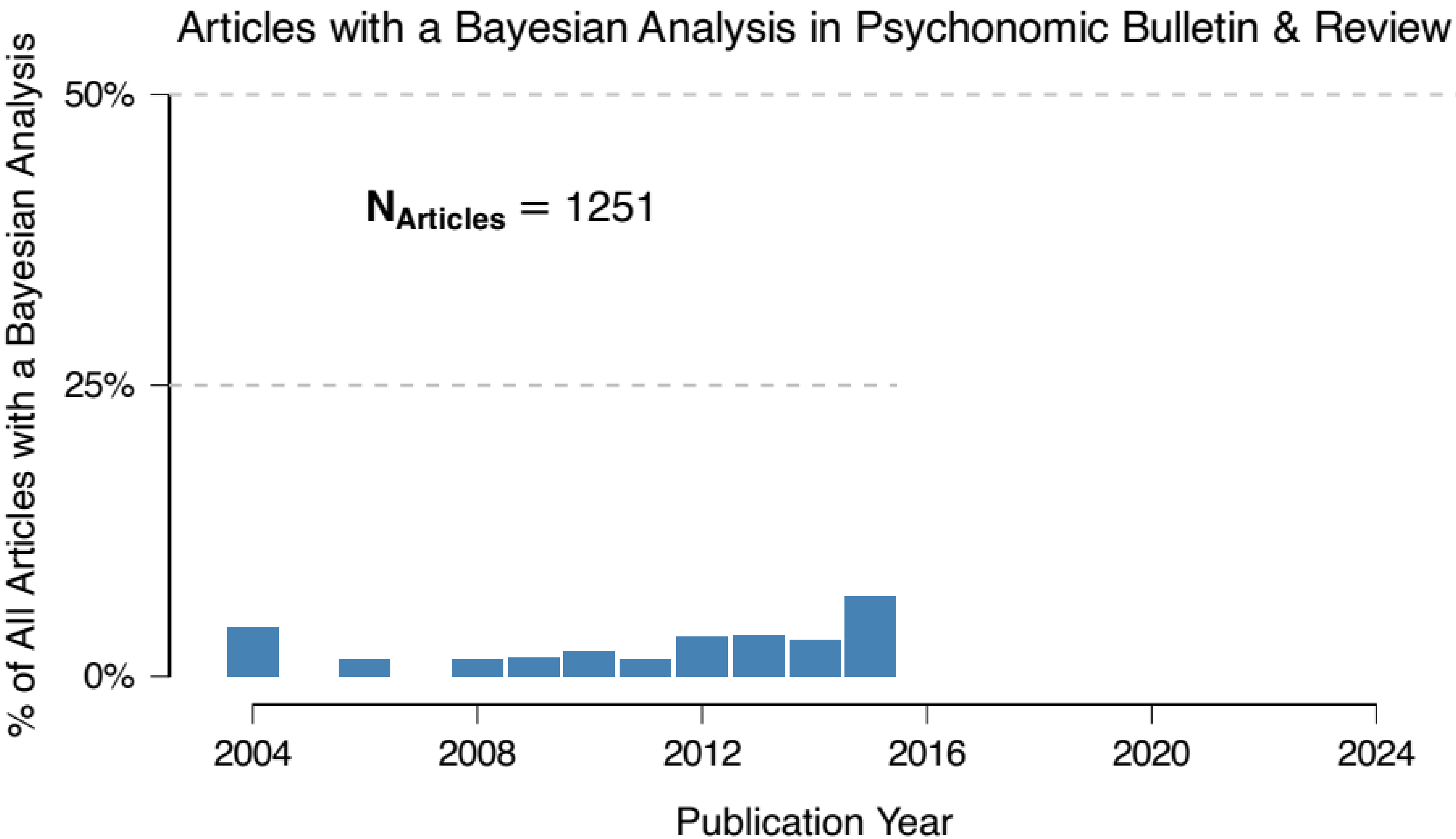
Outline

- What is Bayesian inference?
- Current popularity
- Unique advantages
- Errors: Type B and Type D
- Bayesian hypothesis testing
- Conclusion

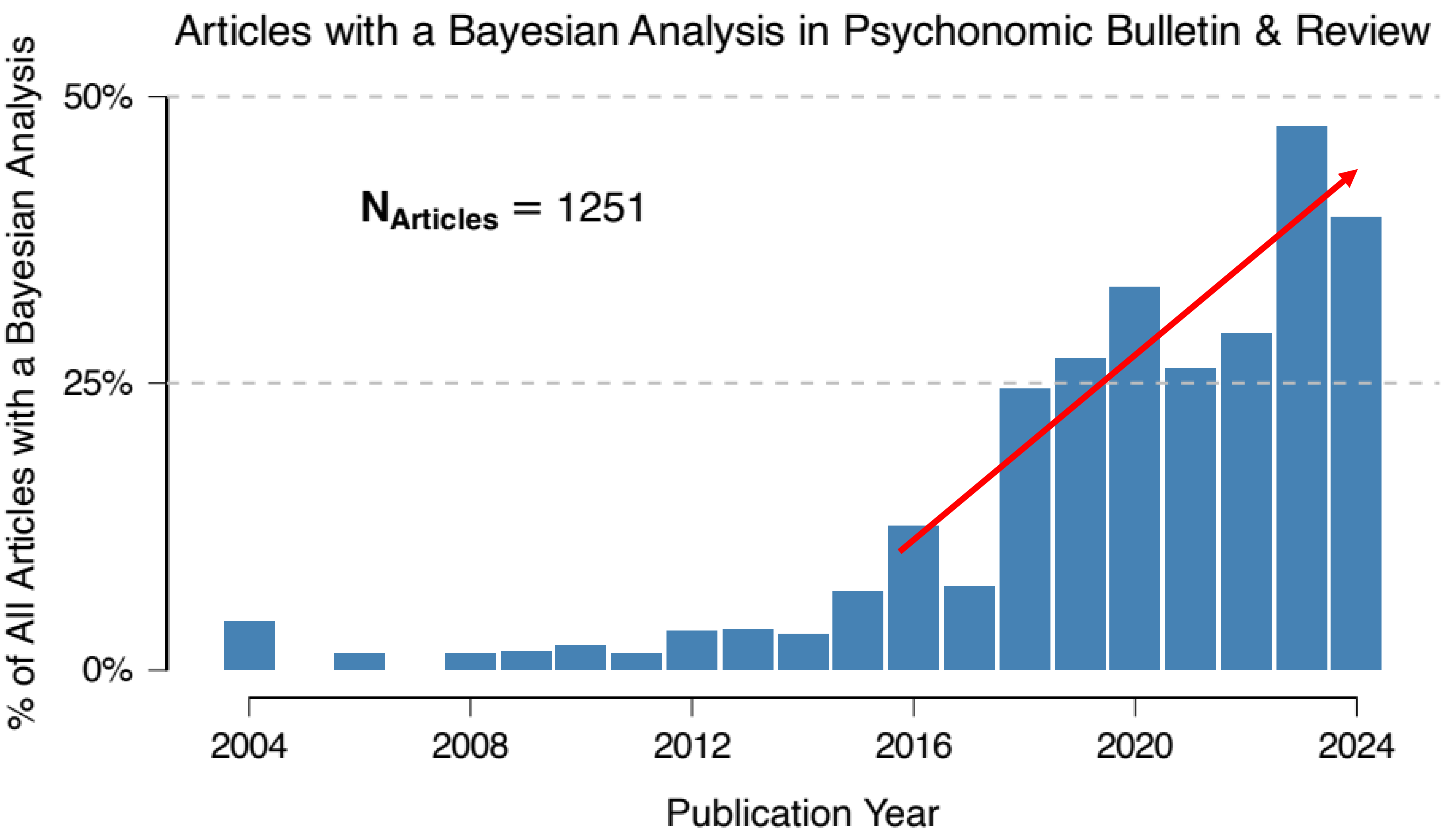
Julius Pfadt
University of Amsterdam

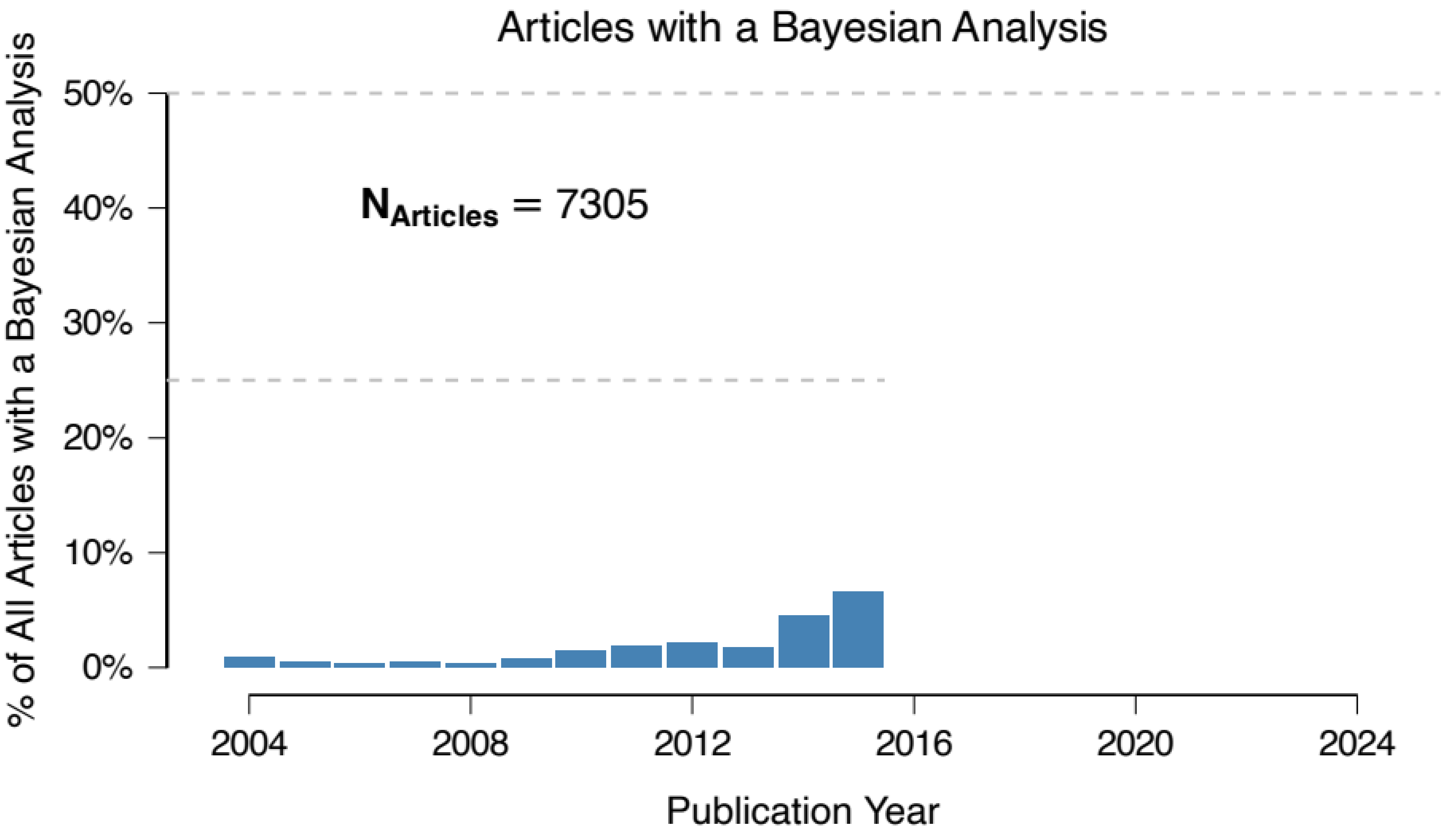



Articles with a Bayesian Analysis in Psychonomic Bulletin & Review
% of All Articles with a Bayesian Analysis
$N_{Articles}$ = 1251
0%
2004
2008
2012
2016
2020
2024
Publication Year

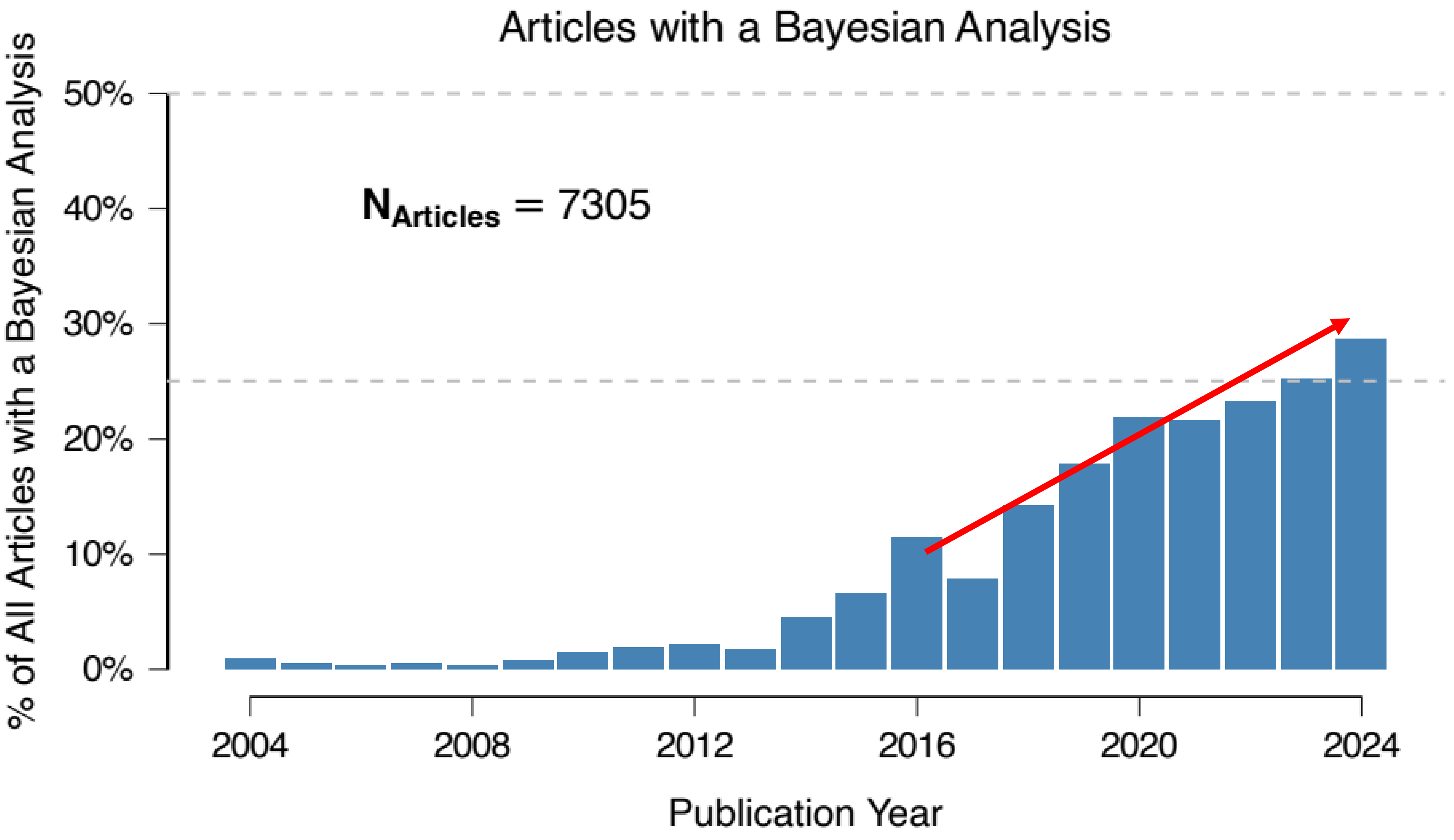


Articles with a Bayesian Analysis in Psychonomic Bulletin & Review
% of All Articles with a Bayesian Analysis
50%
25%
0%
N_Articles = 1251
2004
2008
2012
2016
2020
2024
Publication Year


Articles with a Bayesian Analysis in Psychonomic Bulletin & Review
$N_{Articles}$ = 1251
50%
25%
0%
% of All Articles with a Bayesian Analysis
2004
2008
2012
2016
2020
2024
Publication Year



Articles with a Bayesian Analysis
% of All Articles with a Bayesian Analysis
50%
40%
30%
20%
10%
0%
$N_{Articles}$ = 7305
2004
2008
2012
2016
2020
2024
Publication Year



Articles with a Bayesian Analysis
% of All Articles with a Bayesian Analysis
50%
40%
30%
20%
10%
0%
N_Articles = 7305
2004
2008
2012
2016
2020
2024
Publication Year

JASP

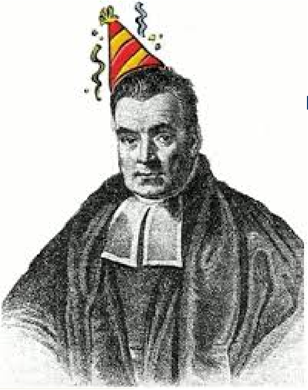
- In order to make Bayesian inference mainstream we have developed JASP, “Jeffreys’s Amazing Statistics Program”.

Harold Jeffreys (1891-1989)

Painting by Marlijn Bouwman

JASP

- JASP is open-source software based on R.
- JASP comes with an attractive graphical user interface.
- JASP allows both Bayesian *and* frequentist analyses.

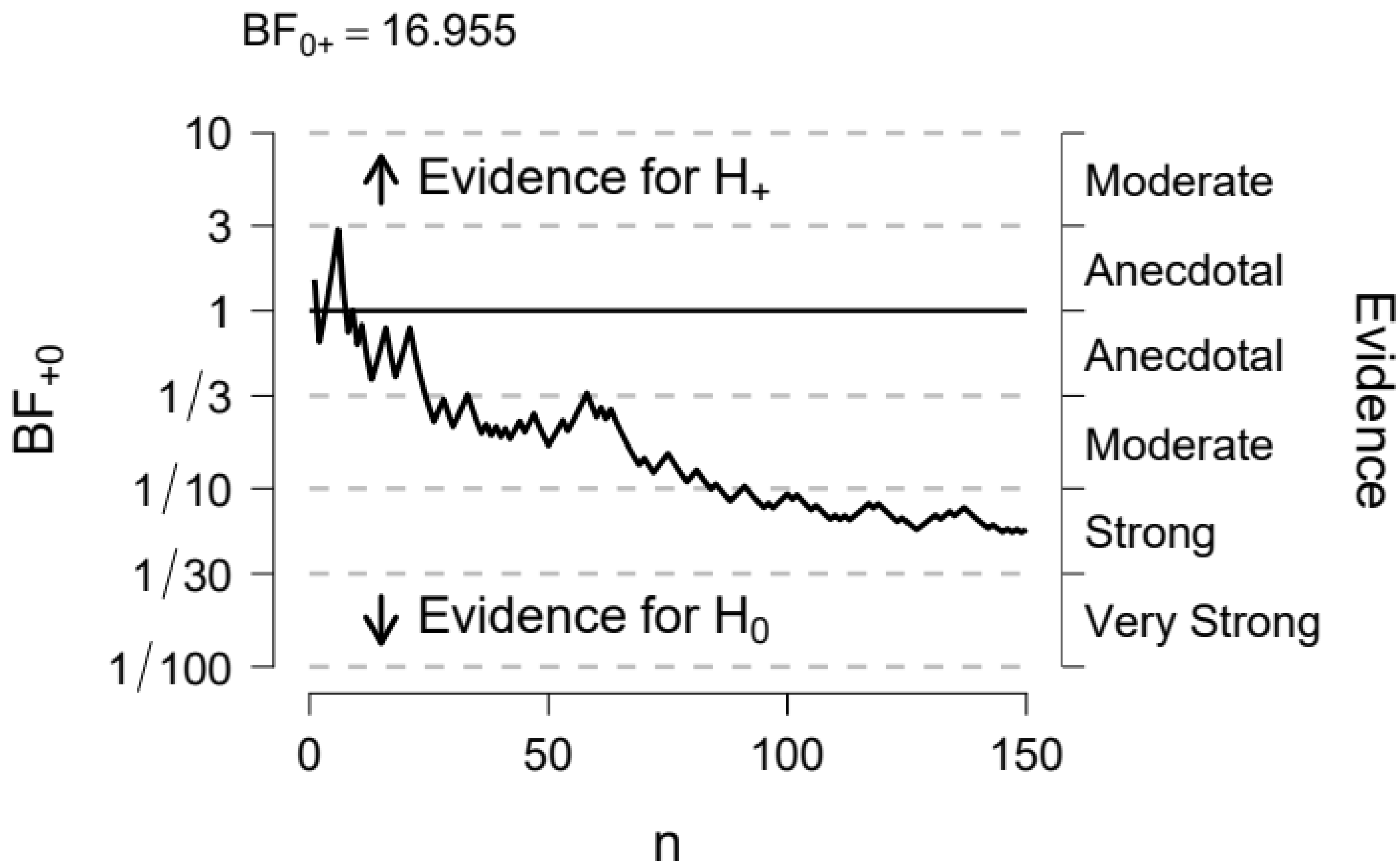


Outline

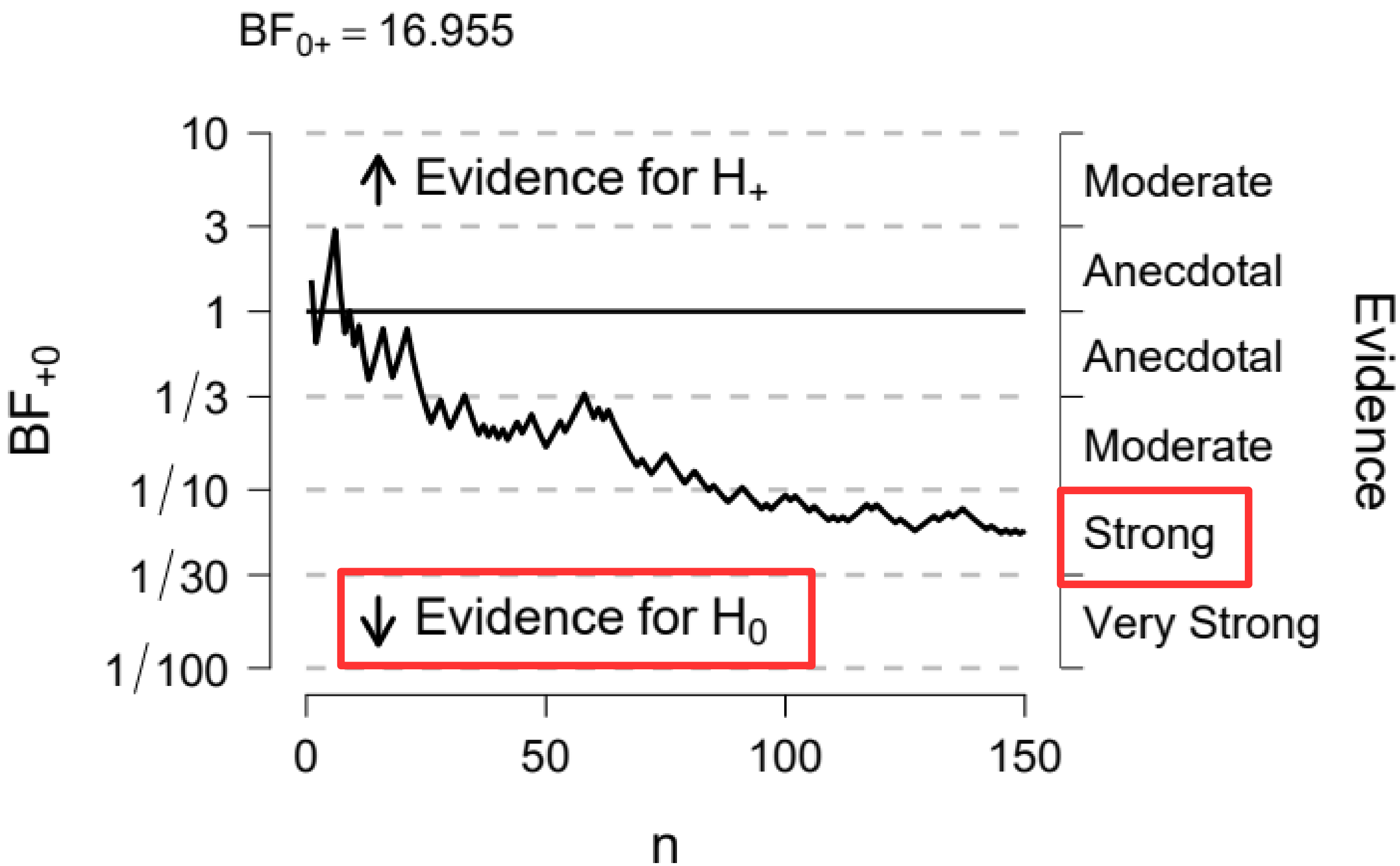
- What is Bayesian inference?
- Current popularity
- Unique advantages
- Errors: Type B and Type D
- Bayesian hypothesis testing
- Conclusion

Example: Children with and without ADHD perform a cognitive task in the fMRI scanner.

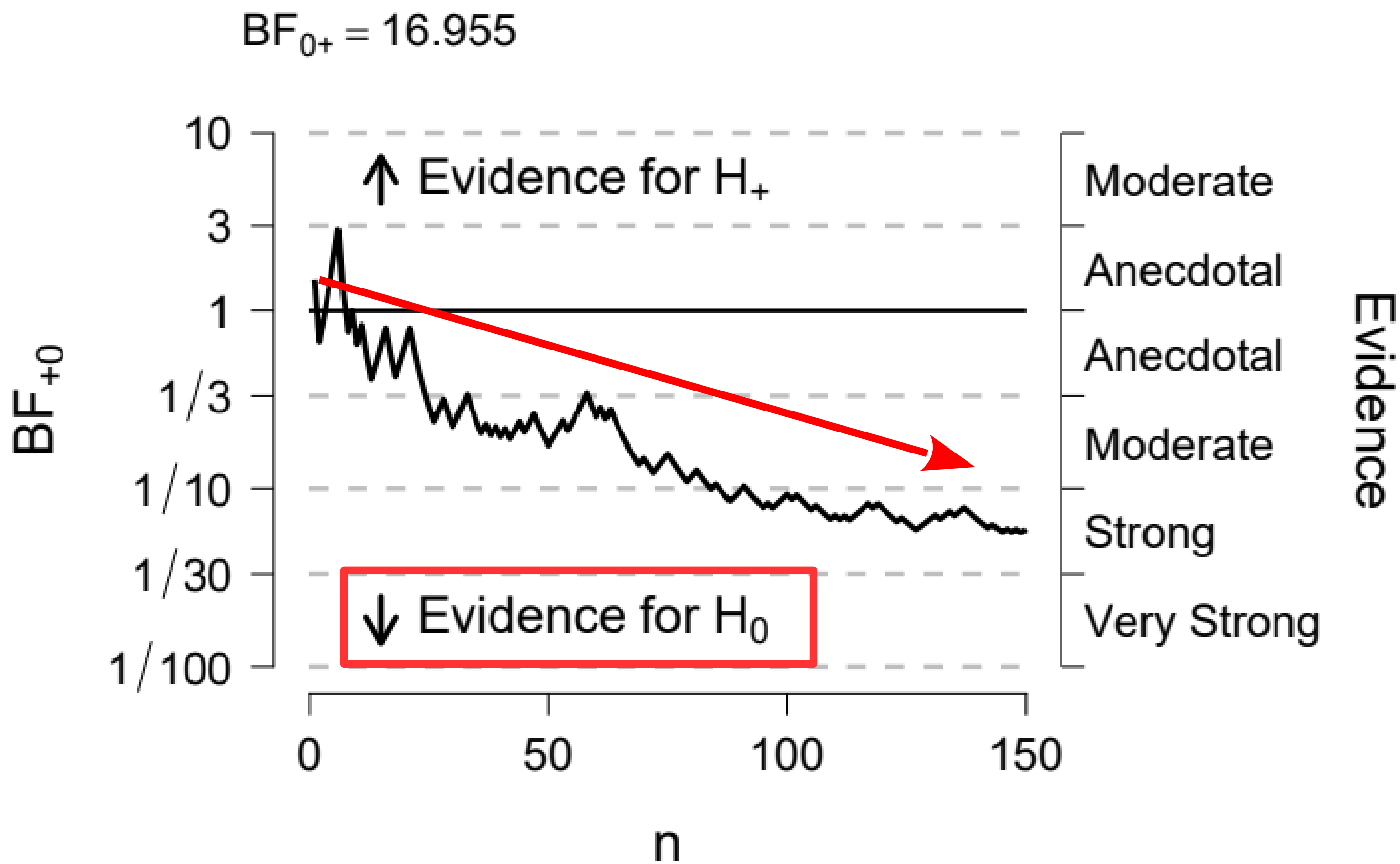
Example: Children with and without ADHD perform a cognitive task in the fMRI scanner. We wish to test the theory that the difference between the two groups is <u>not</u> affected by the surface features of the task.



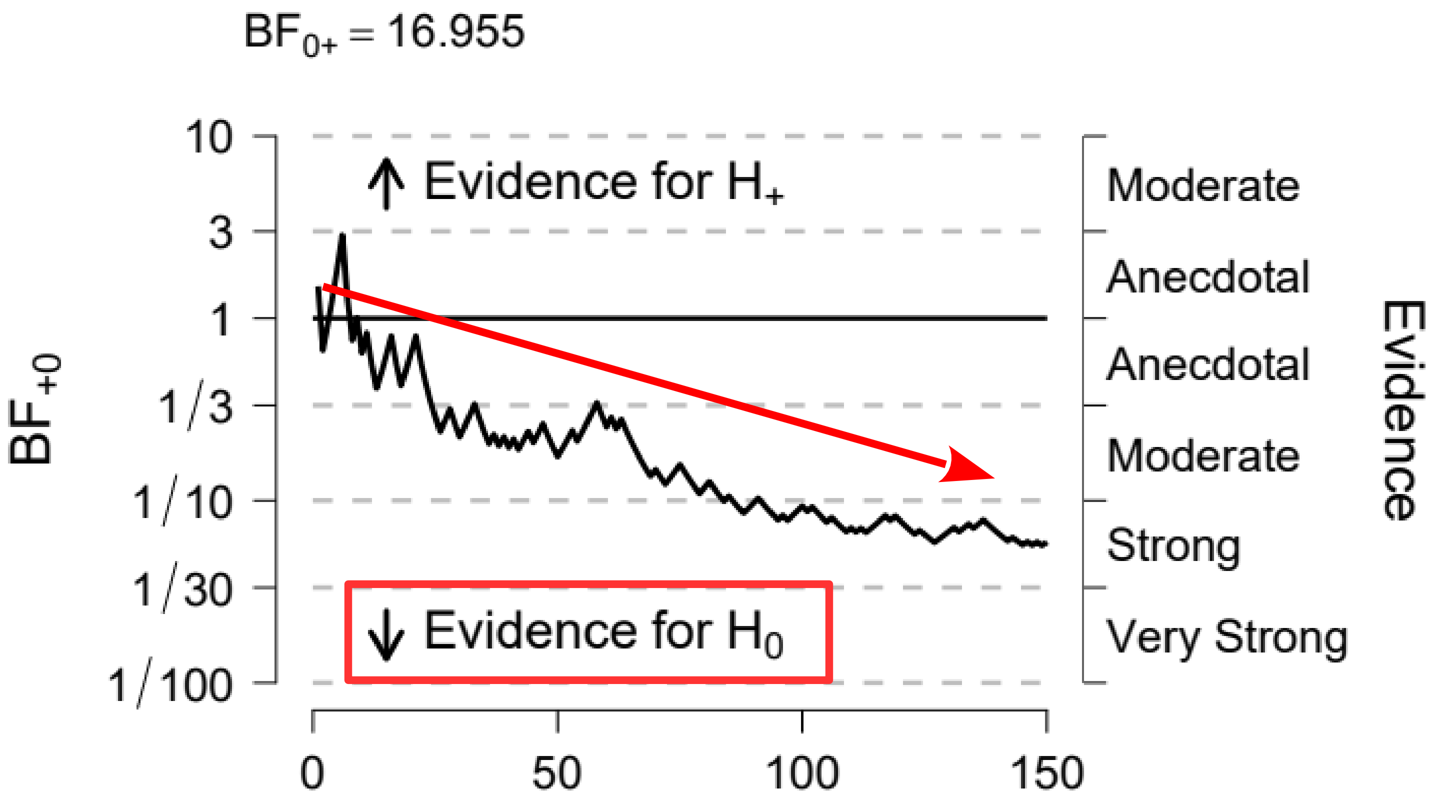
BF0+ = 16.955
10
3
1
1/3
1/10
1/30
1/100
BF+0
Evidence for H+
Evidence for H0
Moderate
Anecdotal
Anecdotal
Moderate
Strong
Very Strong
Evidence
0
50
100
150
n



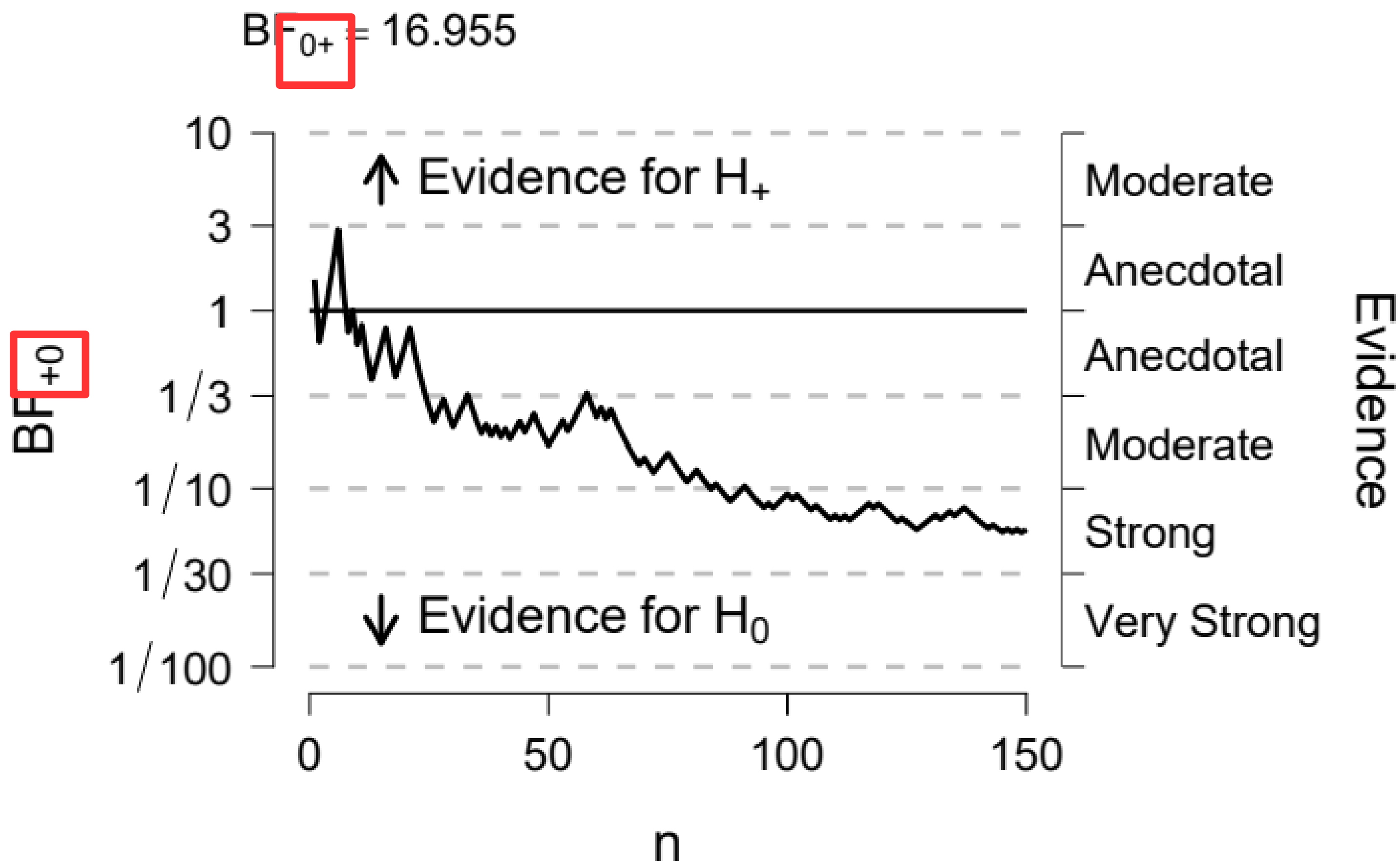
We can have evidence in favor of the *absence* of an effect.



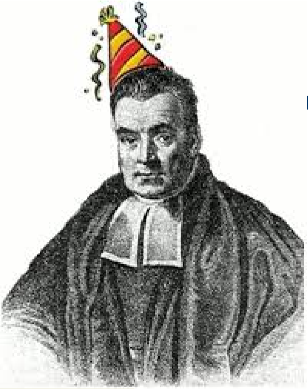
We <u>monitor</u> this evidence as the data accumulate.



This allows evidence-based stopping and continuation, which is efficient and ethical.

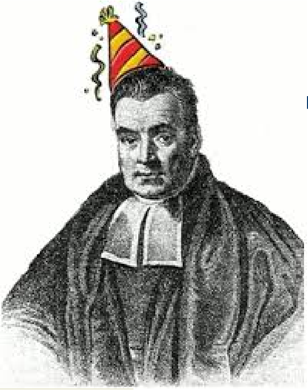


We can <u>incorporate knowledge</u> about the expected direction and size of the effect.



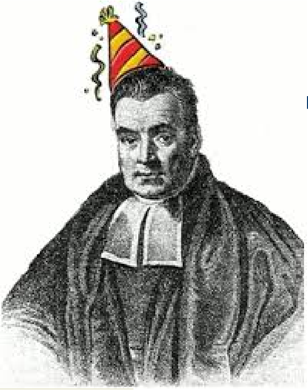
Outline

- What is Bayesian inference?
- Current popularity
- Unique advantages
- Errors: Type B and Type D
- Bayesian hypothesis testing
- Conclusion



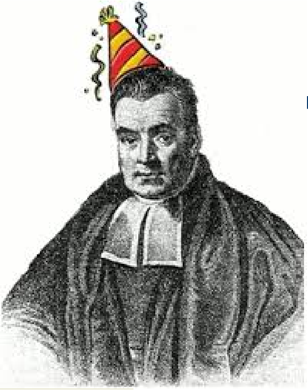
Evidence

- Data can be said to offer *evidence* for a claim when they make that claim more plausible than it was before.



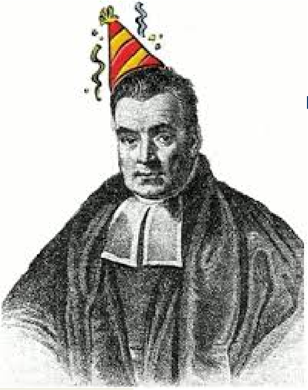
Evidence

- Data can be said to offer *evidence* for a claim when they make that claim more plausible than it was before.
- Hence, evidence is inherently a Bayesian concept, as it refers to a change in credibility.



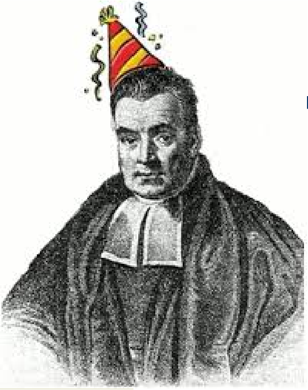
Example

- You publish the claim “Our data show that attention modulates perception of visual space” while arguing that your data make that claim *less* plausible than it was before.



Example

- You publish the claim "Our data show that attention modulates perception of visual space" while arguing that your data make that claim *less* plausible than it was before.
- This would be preposterous.

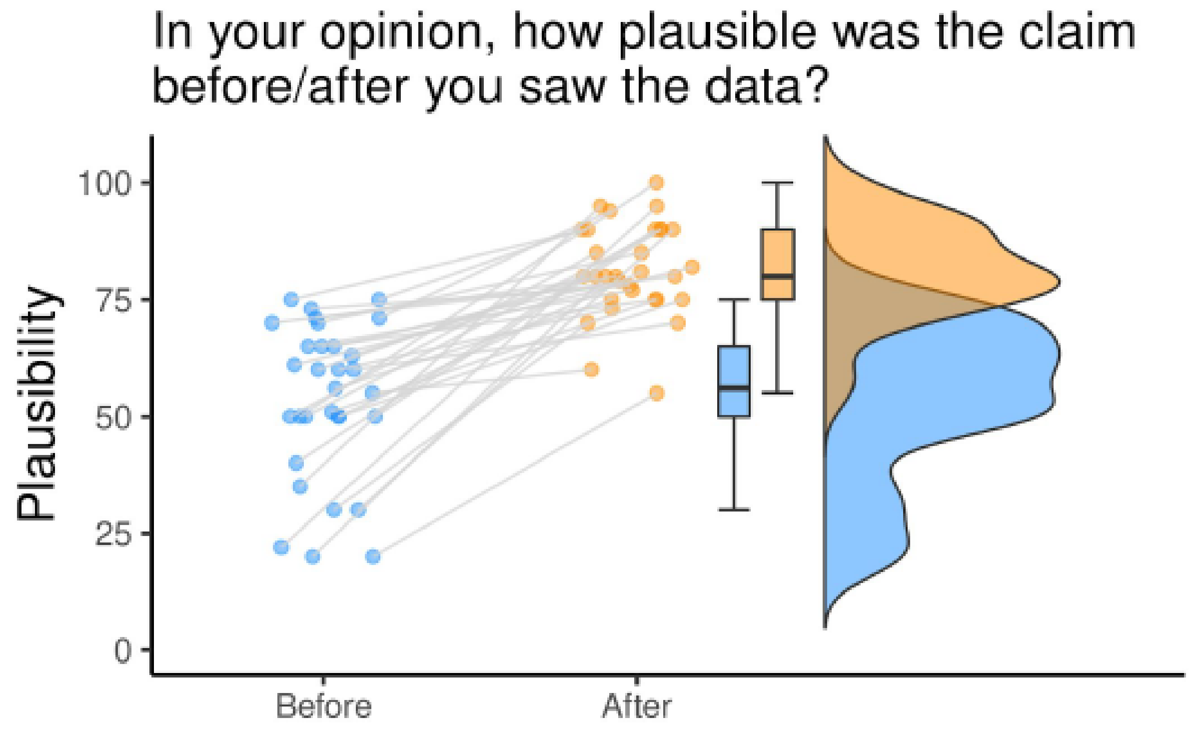


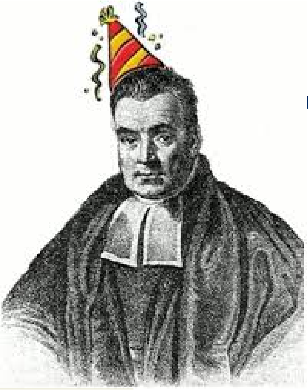
Example

- You publish the claim "Our data show that attention modulates perception of visual space" while arguing that your data make that claim *less* plausible than it was before.
- This would be preposterous.
- Researchers should not find this acceptable, and there is some evidence that they don't.

Strong Public Claims May Not Reflect Researchers' Private Convictions

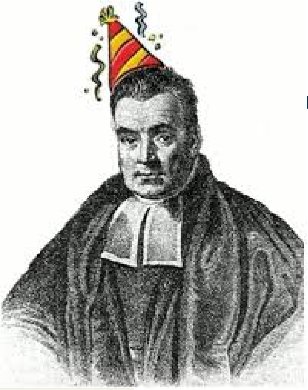
Johnny van Doorn[1], Don van den Bergh[1], Fabian Dablander[1], Noah van Dongen[1], Koen Derks[1], Nathan Evans[2], Quentin Gronau[1], Julia Haaf[1], Yoshihiko Kunisato[3], Alexander Ly[1,4], Maarten Marsman[1], Alexandra Sarafoglou[1], Angelika Stefan[1], Eric-Jan Wagenmakers[1]


In your opinion, how plausible was the claim before/after you saw the data?
Plausibility
100
75
50
25
0
Before
After



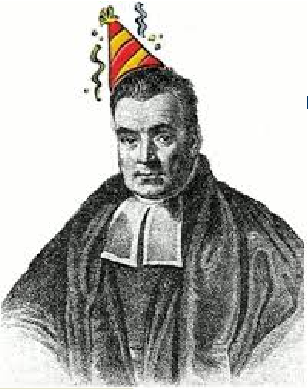
Evidence

- In order to know whether or not we are making a preposterous claim, we need to conduct a Bayesian analysis.



Type B Error

When a reasonable Bayesian analysis undercuts the conclusions from a frequentist analysis.



Type D Error

- Scientific inference is about updating reasonable opinion; it is not about *making decisions*.
- For me, the concept of “making a decision” on a scientific hypothesis makes zero sense.

Rozeboom's Piece of Pie Offered for Dessert

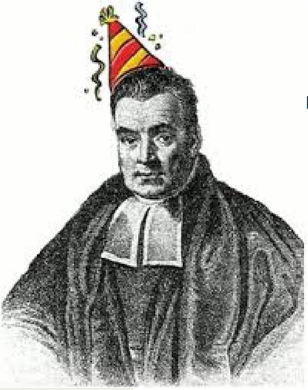
"The null-hypothesis significance test treats 'acceptance' or 'rejection' of a hypothesis as though these were decisions one makes. (...)"

Rozeboom's Piece of Pie Offered for Dessert

"But a hypothesis is not something, like a <u>piece of pie offered for dessert</u>, which can be accepted or rejected by a voluntary physical action.(...)"

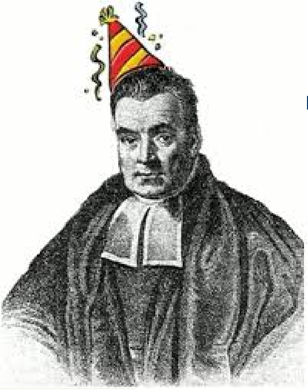
Rozeboom's Piece of Pie Offered for Dessert

"Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true." (Rozeboom, 1960, pp. 422-423)"



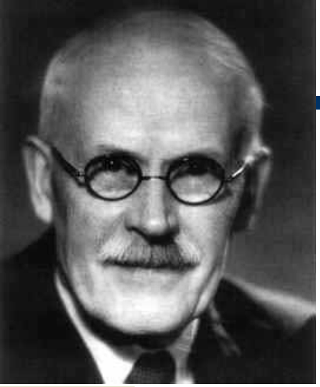
Type D Error

When analysts transmogrify a scientific inference problem into a decision problem.



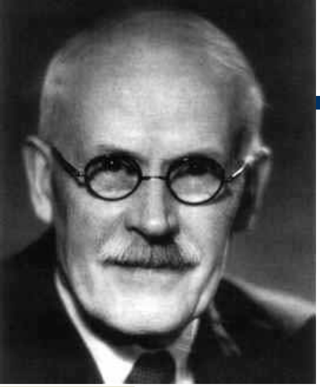
Outline

- What is Bayesian inference?
- Current popularity
- Unique advantages
- Errors: Type B and Type D
- Bayesian hypothesis testing
- Conclusion



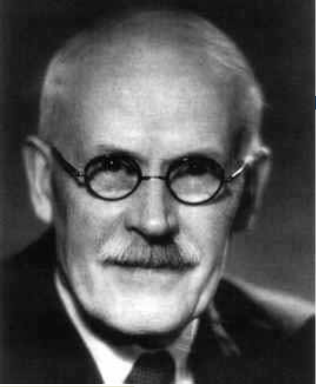
Bayesian Hypothesis Test

- Suppose we have two models, H_0 and H_1.
- Which model is better supported by the data?
- The model that predicted the data best!
- The ratio of predictive performance is known as the Bayes factor (Jeffreys, 1961).



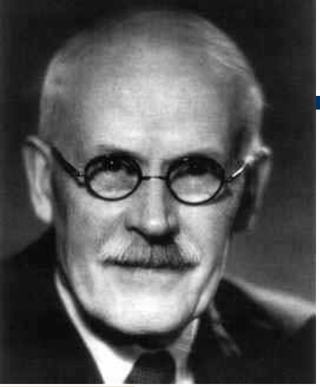
Bayesian Hypothesis Test

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior beliefs} \\ \text{about hypotheses}}}$$



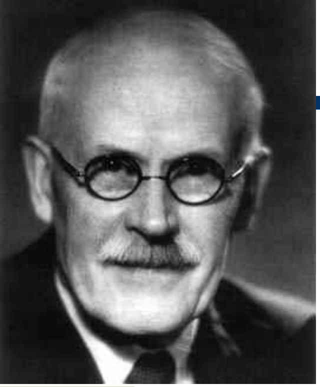
Bayesian Hypothesis Test

$$\underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior beliefs} \\ \text{about hypotheses}}}$$



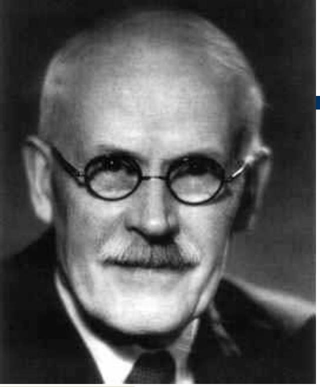
Bayesian Hypothesis Test

$$\underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive} \\ \text{updating factor}}}$$



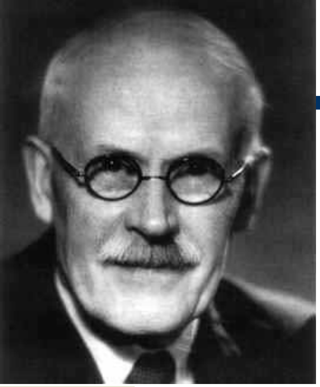
Bayesian Hypothesis Test

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior beliefs} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior beliefs} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive} \\ \text{updating factor}}}$$



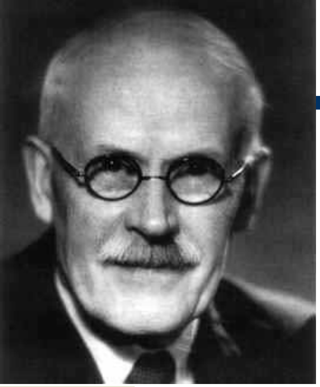
Binomial Hypothesis Test

- Consider the example of pure induction. The null hypothesis (a *universal generalization*) equals "all X are Y".



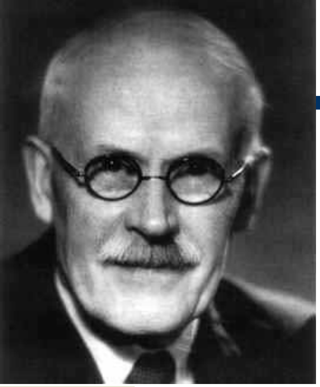
Binomial Hypothesis Test

- Consider the example of pure induction. The null hypothesis (a *universal generalization*) equals "all X are Y".
- You observe only confirmatory instances.



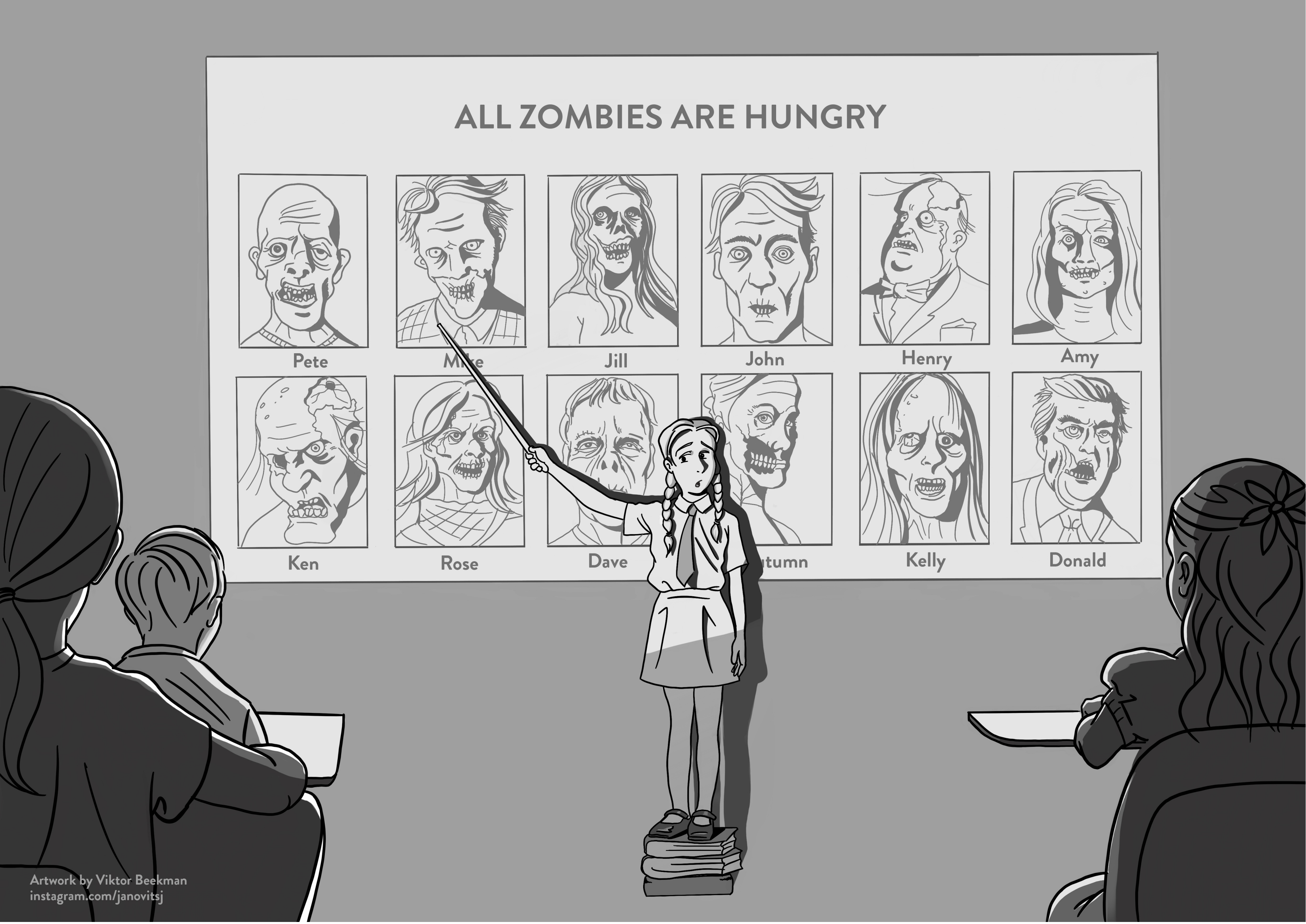
Binomial Hypothesis Test

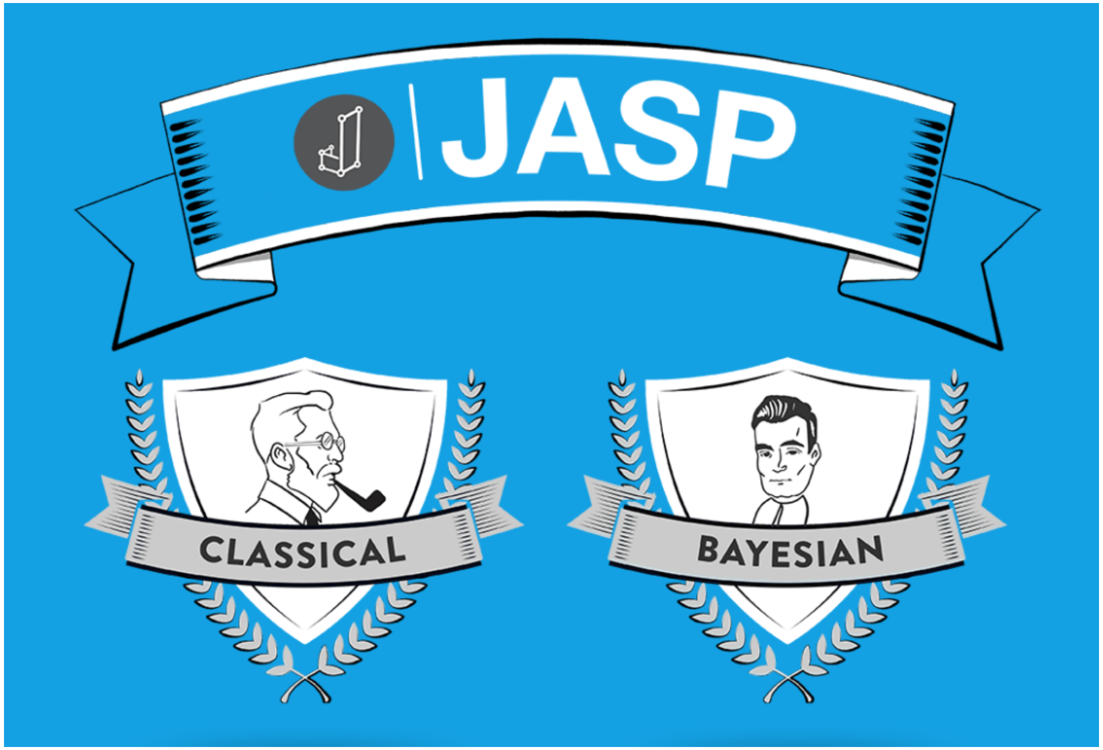
- Consider the example of pure induction. The null hypothesis (a *universal generalization*) equals “all X are Y”.
- You observe only confirmatory instances.
- Every confirmatory instance should increase your confidence in the general law.

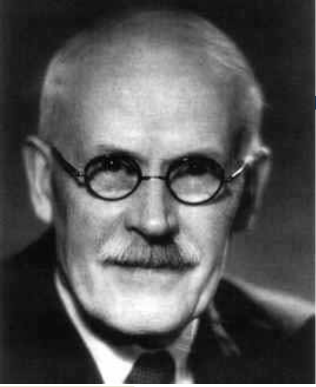


Binomial Hypothesis Test

- Consider the example of pure induction. The null hypothesis (a *universal generalization*) equals “all X are Y”.
- You observe only confirmatory instances.
- Every confirmatory instance should increase your confidence in the general law.
- Let’s see how this works in Bayesian inference.

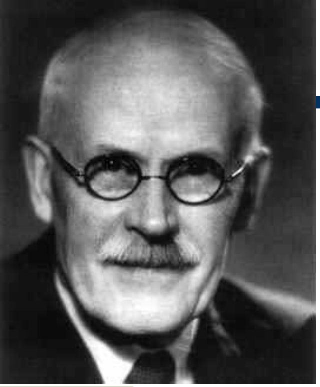

ALL ZOMBIES ARE HUNGRY
Pete
Jill
John
Henry
Amy
Ken
Rose
Dave
Kelly
Donald
Artwork by Viktor Beekman
instagram.com/janovitsj


JASP
CLASSICAL
BAYESIAN



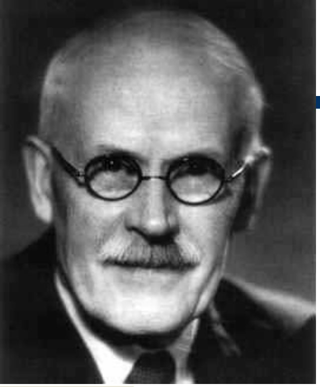
Bayesian Hypothesis Test

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior beliefs} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior beliefs} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive} \\ \text{updating factor}}}$$



Properties of the Bayes factor

- Sensitive to prior information
- Independent of prior model probability
- *Consistent* under H1 and under H0
- *Relative* measure of evidence



Important Aspects

- Bayes factors discriminate between *absence of evidence* and *evidence of absence*.
- Bayes factors may be monitored as the data accumulate.

Concrete Examples

Example I: Fair or Biased?

Is the Coin Fair?

- A coin is flipped and lands “heads” 8 out of 9 times: H, H, H, H, H, H, H, H, T.

Is the Coin Fair?

- A coin is flipped and lands “heads” 8 out of 9 times: H, H, H, H, H, H, H, H, T.
- Do these data provide evidence that the coin is unfair?

Is the Coin Fair?

- A coin is flipped and lands "heads" 8 out of 9 times: H, H, H, H, H, H, H, H, T.
- Do these data provide evidence that the coin is unfair?
- NB. The *p*-value equals .04 ("reject the null hypothesis").

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, $\theta = \frac{1}{2}$.
- H1: the coin is double-heads, $\theta = 1$.

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, $\theta = ½$.
- H1: the coin is double-heads, $\theta = 1$.
- Conclusion: infinite evidence **in favor of** the fair coin!

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, $\theta = ½$.
- H1: the coin is very slightly biased, $\theta = 0.51$.

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, θ = ½.
- H1: the coin is very slightly biased, θ = 0.51.
- Conclusion: BF10 = 1.15, almost **no evidence** at all.

Who Won?

Drawing by Dirk-Jan Hoek

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, $\theta = \frac{1}{2}$.
- H1: $\theta \sim$ uniform(0,1) [“anything goes”].

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, $\theta = ½$.
- H1: $\theta \sim$ uniform(0,1) [“anything goes”].
- Conclusion: BF10 = 5.67, **moderate evidence** against H0.

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, $\theta = ½$.
- H1: $\theta \sim$ beta(10,10) [“θ is near ½”].

Is the Coin Fair? HHHHHHHHT

- H0: the coin is fair, $\theta = ½$.
- H1: $\theta \sim$ beta(10,10) [“θ is near ½”].
- Conclusion: BF10 = 2.00, **weak evidence** against H0.

Is the Coin Fair? HHHHHHHHT

- The reason we obtained different answers is because we were asking different questions!

Before you give an answer, consider the question

Jeffreys's platitude

Madness

- Some people do not like Bayes factors.

Madness

- Some people do not like Bayes factors.
- They would like very *different questions* to result in very *similar answers*.

Madness

- Some people do not like Bayes factors.
- They would like very *different questions* to result in very *similar answers*.
- In other words, they would prefer a method that is *less sensitive to the prior distribution*.

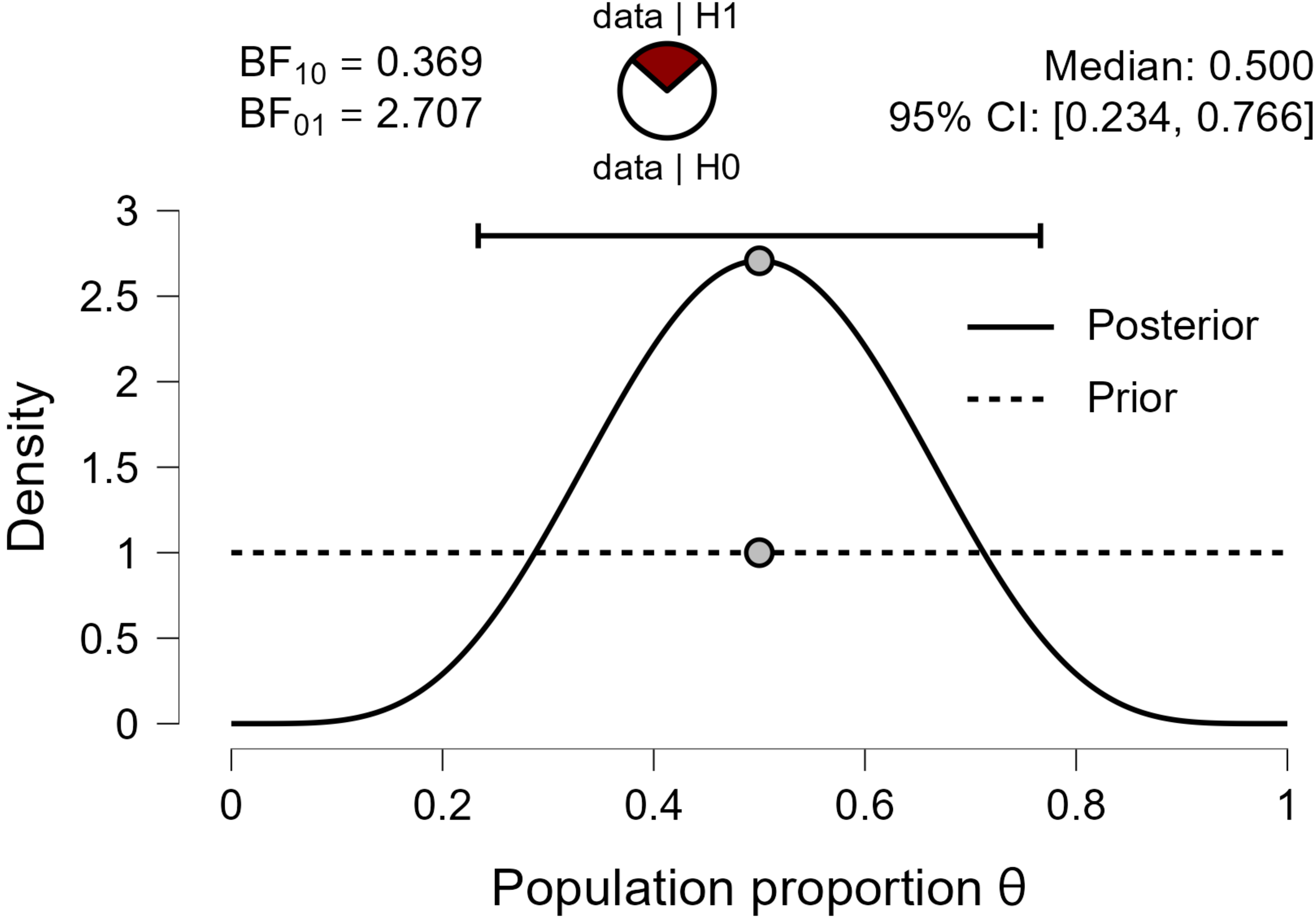
Madness

- Some peopl… …ayes factors.
- They would … …*rent questions* to result in very… …*ers*.
- In other wor… …l prefer a method that is *less se*… …*prior distribution*.

Therein lies <u>madness</u>!

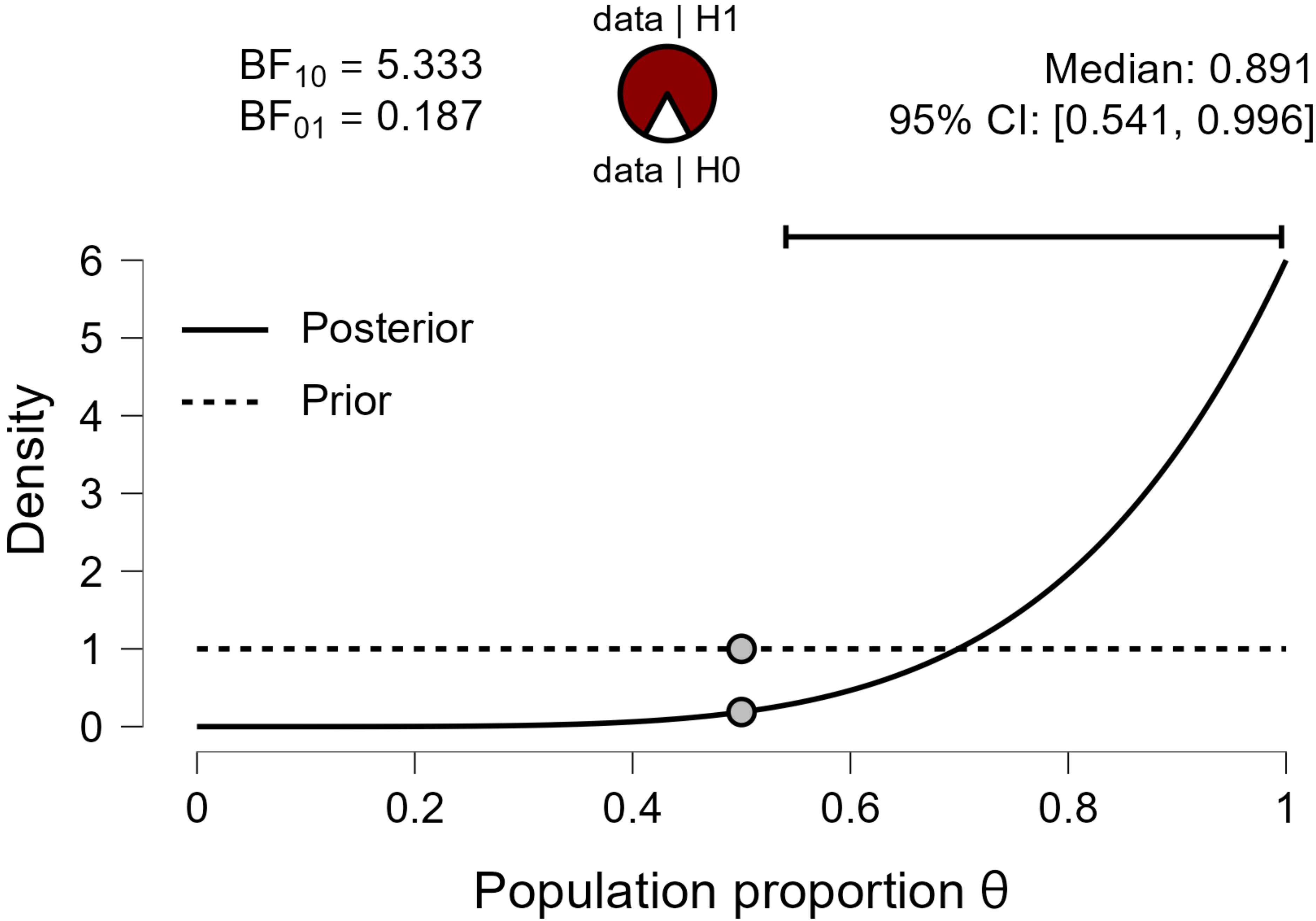
Dynamic Coherence

- We test H0: $\theta = ½$ versus H1: $\theta \sim \text{beta}(1,1)$
- [both specifications may be generalized]
- We find that $s = f = n/2$: an equal split, and this has to be evidence in favor of H0.
- Consider the example of $s = f = 5$...


data | H1
BF10 = 0.369
BF01 = 2.707
Median: 0.500
95% CI: [0.234, 0.766]
data | H0
Density
3
2.5
2
1.5
1
0.5
0
Posterior
Prior
0
0.2
0.4
0.6
0.8
1
Population proportion θ

Dynamic Coherence

- Now we split the data into *two* batches.
- It does not matter how we split, but for clarity the first batch has 5 successes and 0 failures, whereas the second batch has 0 successes and 5 failures.
- If we update batch-by-batch we ought to retrieve the original result (coherence!).



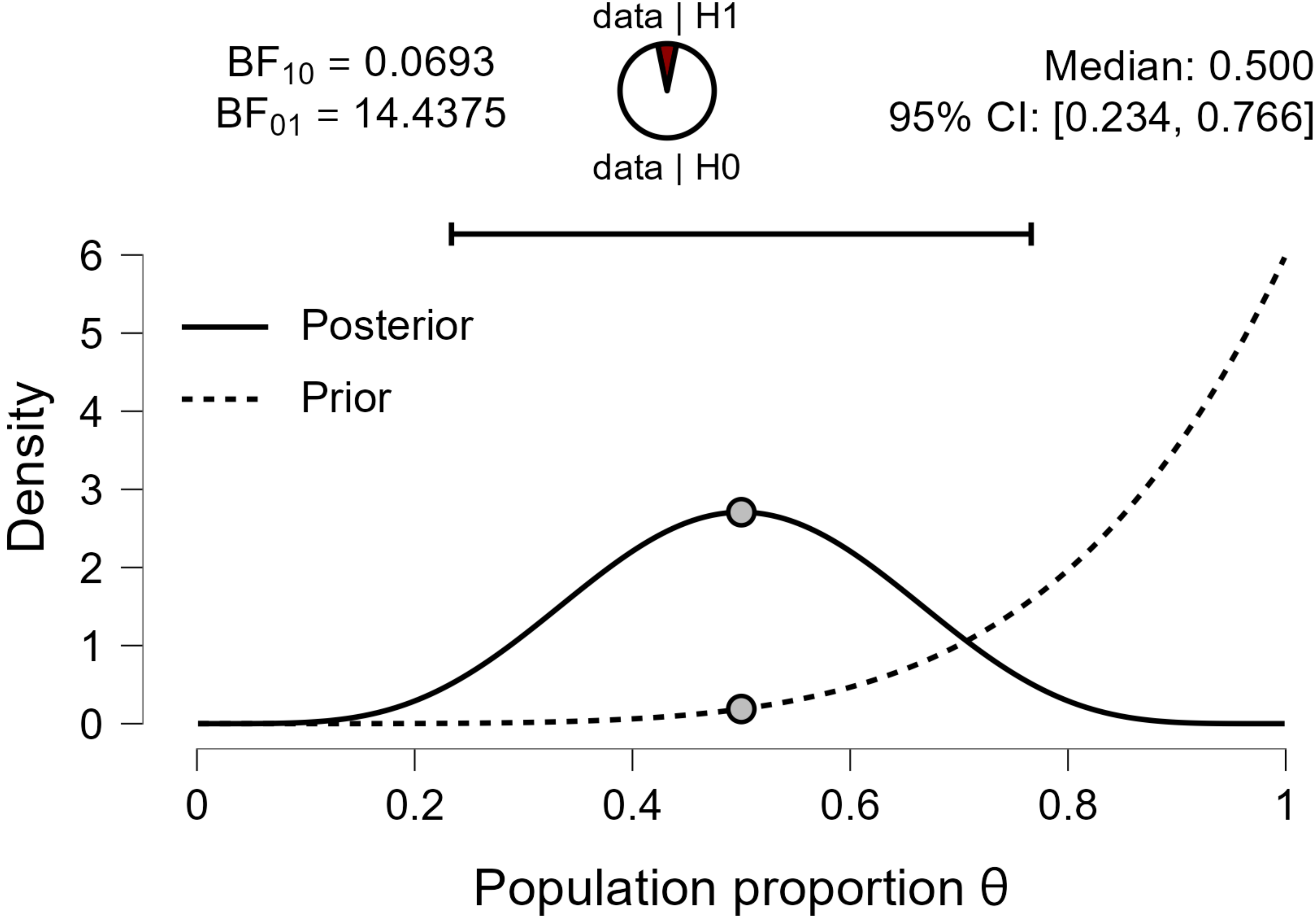
BF_{10} = 5.333
BF_{01} = 0.187
data | H1
data | H0
Median: 0.891
95% CI: [0.541, 0.996]
Posterior
Prior
Density
0
1
2
3
4
5
6
0
0.2
0.4
0.6
0.8
1
Population proportion θ

Dynamic Coherence

- So batch A favors H1.
- But we know the complete data favors H0.

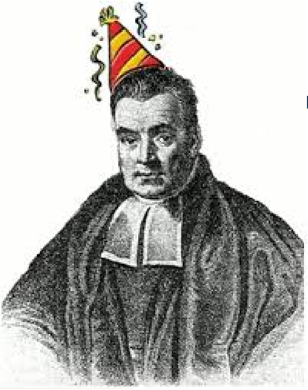
Dynamic Coherence

- So batch A favors H1.
- But we know the complete data favors H0.
- Hence, batch B must favor H0. Also, the strength of this evidence should be higher than what batch A provided for H1.

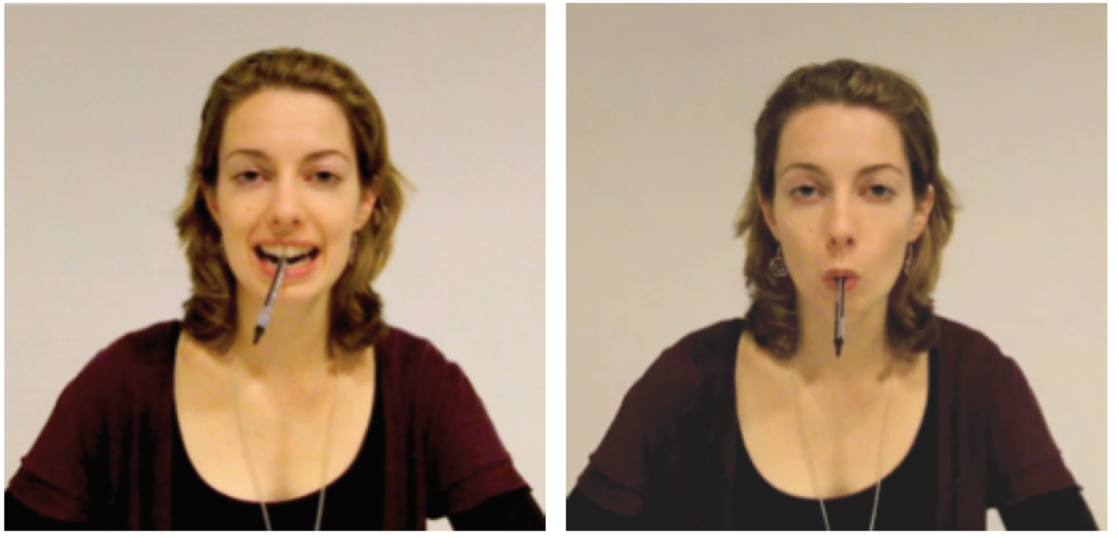

data | H1
BF10 = 0.0693
BF01 = 14.4375
Median: 0.500
95% CI: [0.234, 0.766]
data | H0
Posterior
Prior
Density
6
5
4
3
2
1
0
0
0.2
0.4
0.6
0.8
1
Population proportion θ

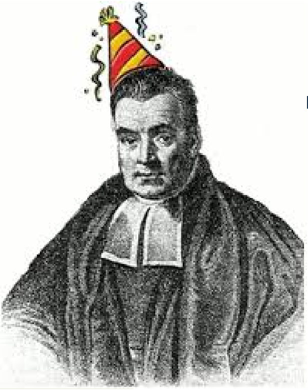
Dynamic Coherence

- What is needed for coherence:
  - The ability to *strongly* prefer H0 over H1;
  - A *unique dependence on the prior distribution!* Batch A pushes θ in the wrong direction, so that the data from batch B are relatively surprising.



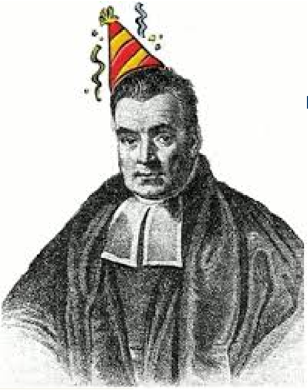
Example II: The Facial Feedback Hypothesis





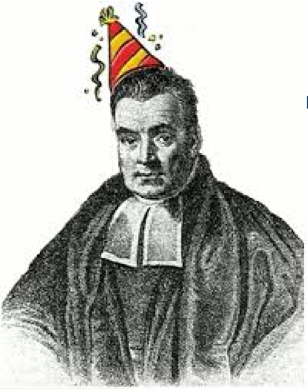
The t-Test

- Main question: “is there an effect?”
  – Skeptic’s H0: there is no effect
  – Proponents’s H1: there is an effect



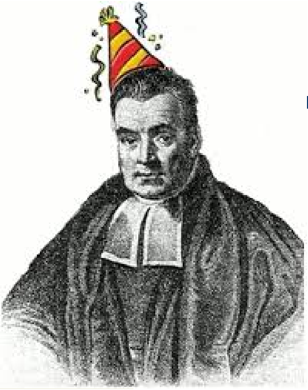
Choosing the Prior: "Subjective" Approach

- The literature suggests the kinds of effect sizes that are plausible;
- Earlier experiments on similar topics may give more specific information;
- Expert knowledge yields relatively precise predictions;
- Drawback: effortful and "subjective".



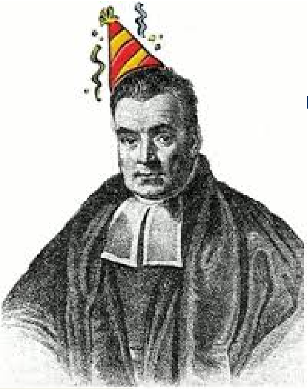
Subjective Prior Distributions On Effect Size

- Prior elicitation with Dr. Suzanne Oosterwijk:
  - H1: $\delta \sim t(\text{mean} = .35, \text{sd} = .102, \text{df} = 3)$
  - δ only allowed to be positive



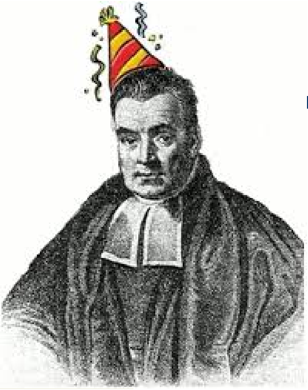
Subjective Prior Distributions On Effect Size

- Prior elicitation (for a different phenomenon, but also small to medium effect) with Dr. Kathleen Vohs:
  - H1: $\delta \sim N(\text{mean} = .30, \text{sd} = .15)$
  - δ only allowed to be positive



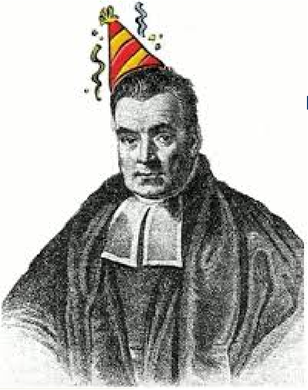
Subjective Prior Distributions On Effect Size

- So we can apply to the data:
  - Oosterwijk prior
  - Vohs prior
  - Default one-sided Cauchy prior



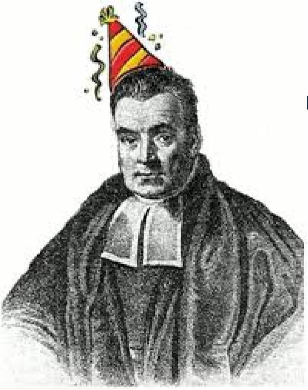
Subjective Prior Distributions On Effect Size

- Use the JASP Summary Stats module.
- Results for Oosterwijk facial feedback experiment:
  - N_smile = 53; N_pout = 57; t = -0.90.



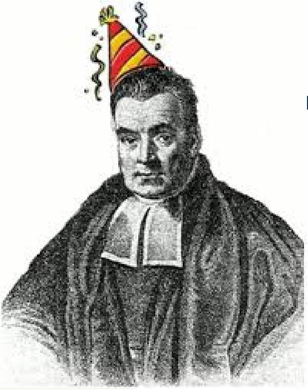
Outline

- What is Bayesian inference?
- Current popularity
- Unique advantages
- Errors: Type B and Type D
- Bayesian hypothesis testing
- Conclusion



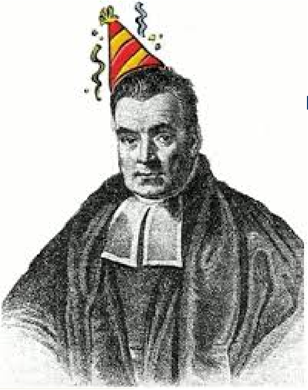
Keeping an Open Mind

- I will support any non-Bayesian method of inference just as long as its meets two modest desiderata:



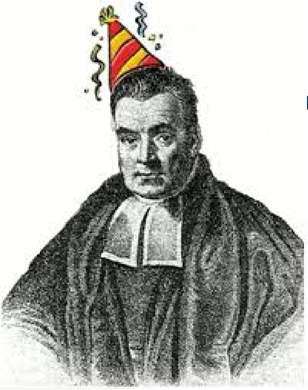
Keeping an Open Mind

- I will support any non-Bayesian method of inference just as long as its meets two modest desiderata:
  - It has to quantify *evidence* in the usual sense of the word (i.e., the change in credibility brought about by the data).



Keeping an Open Mind

- I will support any non-Bayesian method of inference just as long as its meets two modest desiderata:
  - It has to quantify *evidence* in the usual sense of the word (i.e., the change in credibility brought about by the data).
  - It has to be *dynamically coherent*.



Keeping an Open Mind

- I will su esian method of inferenc meets two modest desidera
  - It has *ce* in the usual sense ne change in credil t by the data).
  - It has *coherent*.

Inside every Non-Bayesian, there is a Bayesian struggling to get out

Dennis Lindley