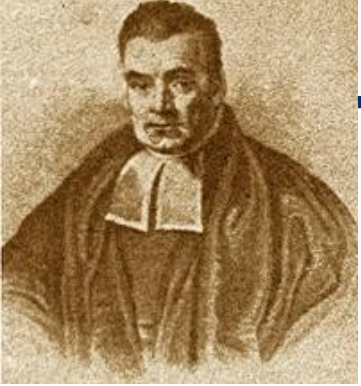


Interim Design Analysis Using Bayes Factor Forecasts

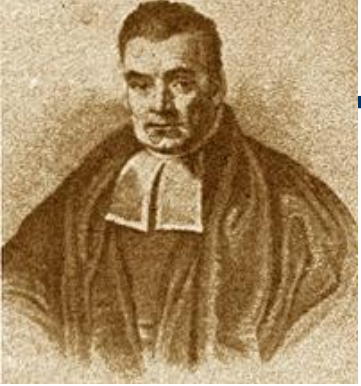


Angelika M. Stefan, Quentin F. Gronau,
& E.-J. Wagenmakers



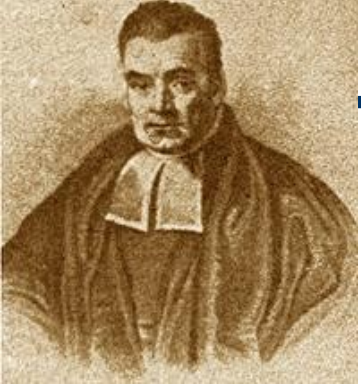
Overview

- ◆ Arguments pro Bayes
- ◆ Stubborn and wrong: when Bayes fails
- ◆ When frequentists are stubborn and wrong
- ◆ Bayes factors
- ◆ Interim design analyses



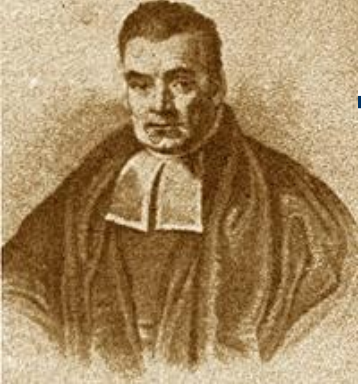
Overview

- ◆ Arguments pro Bayes
- ◆ Stubborn and wrong: when Bayes fails
- ◆ When frequentists are stubborn and wrong
- ◆ Bayes factors
- ◆ Interim design analyses



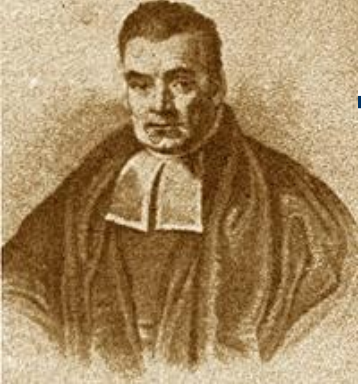
Arguments Pro Bayes

- ◆ Coherent (avoids internal contradictions)



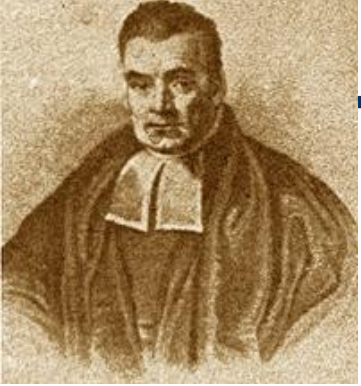
Arguments Pro Bayes

- ◆ Coherent (avoids internal contradictions)
- ◆ Consistent under the null



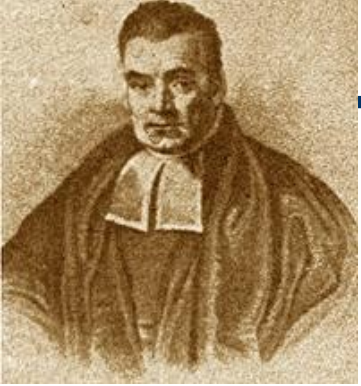
Arguments Pro Bayes

- ◆ Coherent (avoids internal contradictions)
- ◆ Consistent under the null
- ◆ Quantifies evidence (data-driven change in reasonable belief)



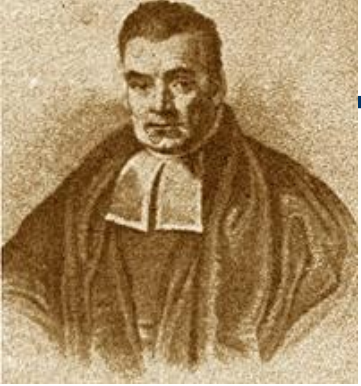
Arguments Pro Bayes

- ◆ Coherent (avoids internal contradictions)
- ◆ Consistent under the null
- ◆ Quantifies evidence (data-driven change in reasonable belief)
- ◆ Adheres to the likelihood principle and stopping rule principle



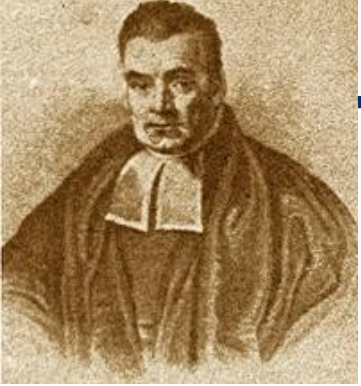
Arguments Pro Bayes

- ◆ Coherent (avoids internal contradictions)
- ◆ Consistent under the null
- ◆ Quantifies evidence (data-driven change in reasonable belief)
- ◆ Adheres to the likelihood principle and stopping rule principle
- ◆ Conditions on the observed data



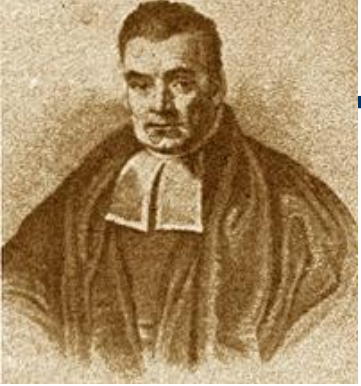
Arguments Pro Bayes

- ◆ Evidence may be monitored as the data accumulate



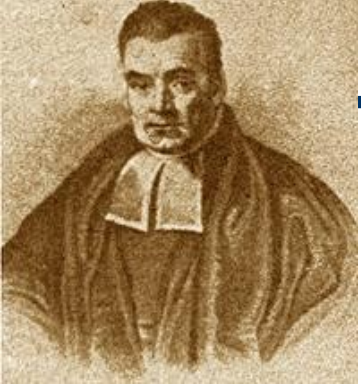
Arguments Pro Bayes

- ◆ Evidence may be monitored as the data accumulate
- ◆ Background knowledge can be incorporated



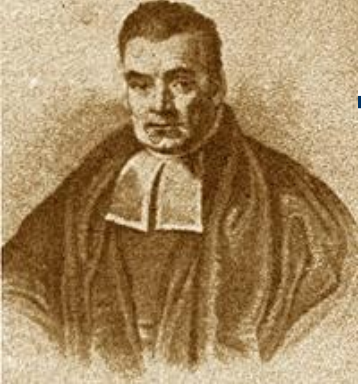
Arguments Pro Bayes

- ◆ Evidence may be monitored as the data accumulate
- ◆ Background knowledge can be incorporated
- ◆ Differentiates evidence for absence from absence of evidence



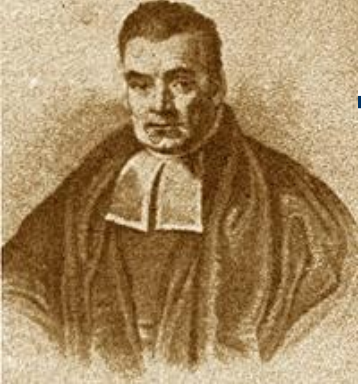
Arguments Pro Bayes

- ◆ Evidence may be monitored as the data accumulate
- ◆ Background knowledge can be incorporated
- ◆ Differentiates evidence for absence from absence of evidence
- ◆ Allows probability statements for parameters and hypotheses



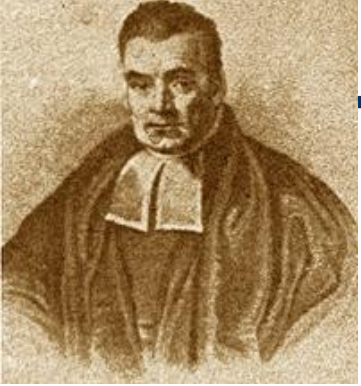
Arguments Pro Bayes

- ◆ Evidence may be monitored as the data accumulate
- ◆ Background knowledge can be incorporated
- ◆ Differentiates evidence for absence from absence of evidence
- ◆ Allows probability statements for parameters and hypotheses
- ◆ Directly addresses the questions of interest



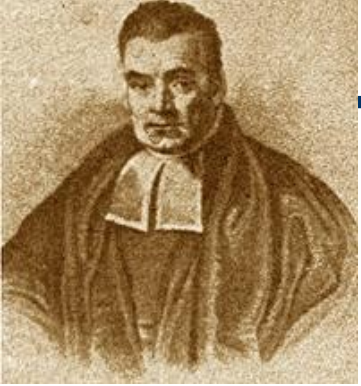
Arguments Pro Bayes

- ◆ Why do regulatory bodies stick to frequentist statistics?



Arguments Pro Bayes

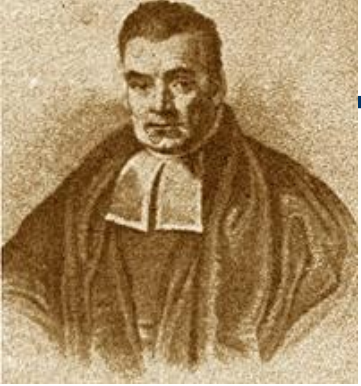
- ◆ Why do regulatory bodies stick to frequentist statistics?
 - *Tradition/inertia* (convenience, fear of the unknown)



Arguments Pro Bayes

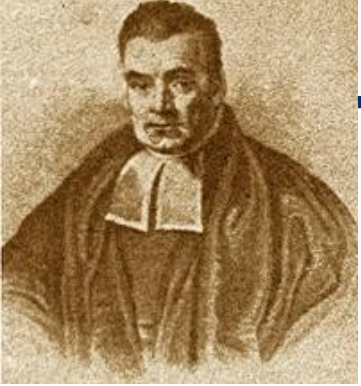
- ◆ Why do regulatory bodies stick to frequentist statistics?
 - *Tradition/inertia* (convenience, fear of the unknown)





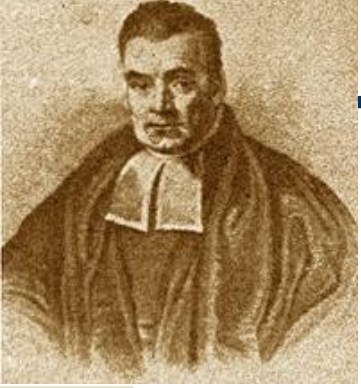
Arguments Pro Bayes

- ◆ Why do regulatory bodies stick to frequentist statistics?



Arguments Pro Bayes

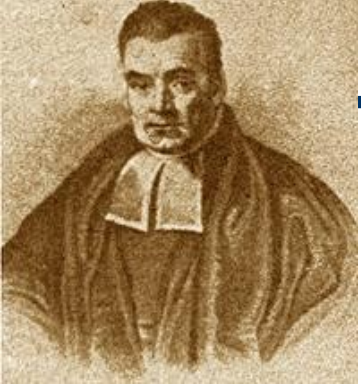
- ◆ Why do regulatory bodies stick to frequentist statistics?
 - “*Laziness*” (Harold Jeffreys)



Arguments Pro Bayes

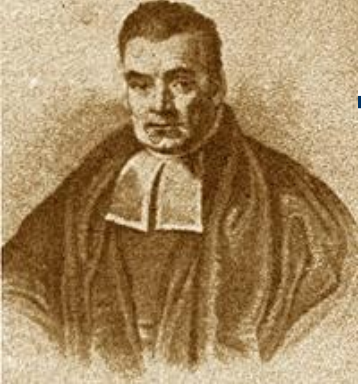
- ◆ Why do regulatory bodies stick to frequentist statistics?
 - “*Laziness*” (Harold Jeffreys)





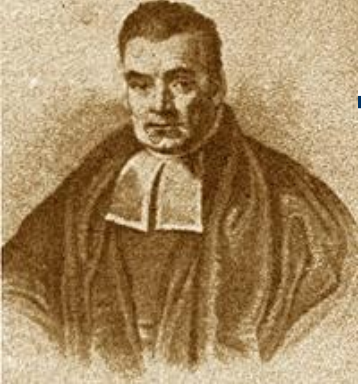
Arguments Pro Bayes

- ◆ A regulatory body is about to allow a drug on the market, based exclusively on a frequentist analysis.



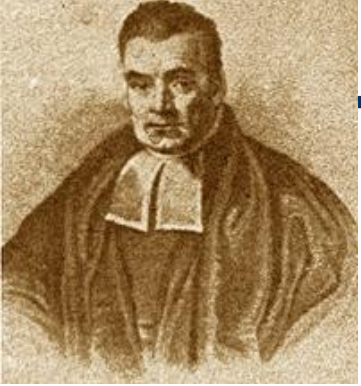
Arguments Pro Bayes

- ◆ A regulatory body is about to allow a drug on the market, based exclusively on a frequentist analysis.
- ◆ Suppose a reasonable Bayesian re-analysis undercuts the frequentist conclusions.



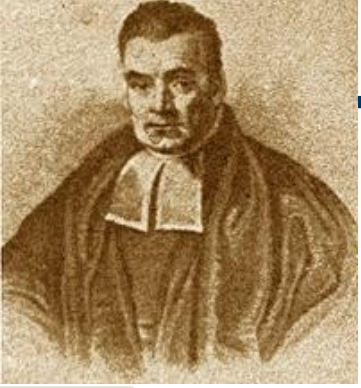
Arguments Pro Bayes

- ◆ A regulatory body is about to allow a drug on the market, based exclusively on a frequentist analysis.
- ◆ Suppose a reasonable Bayesian re-analysis undercuts the frequentist conclusions.
- ◆ Should the regulatory body be aware? Should they care? Should they prevent this possibility from arising?



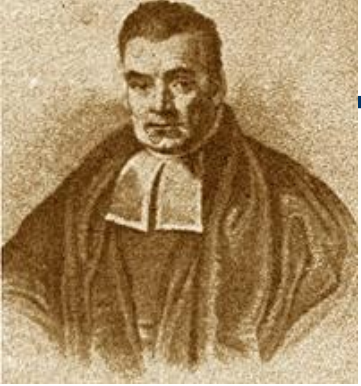
Type B Error

- ◆ Arises when a reasonable Bayesian analysis yields a conclusion that conflicts with the frequentist analysis.
- ◆ Currently, this error is entirely ignored.



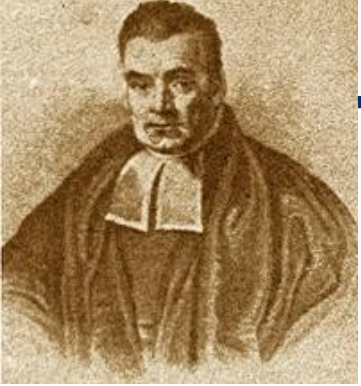
Type B Error

- ◆ Controlling Type I error rate is commendable, but not at the expense of:
 - Quantifying the evidence
 - Assessing the probability that you are correct *for the case at hand*.



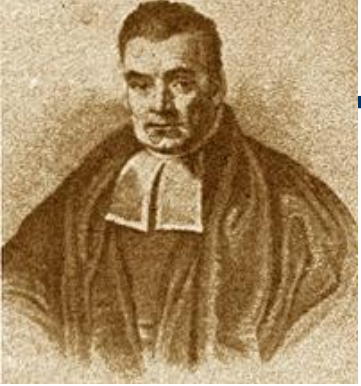
Type F Error

- ◆ We may also introduce the “Type F” Error: executing a frequentist analysis to answer questions that are fundamentally Bayesian. However, I don’t want to be too provocative.



Overview

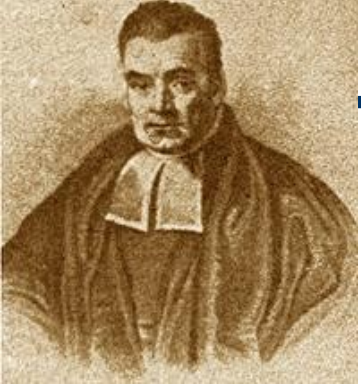
- ◆ Arguments pro Bayes
- ◆ **Stubborn and wrong: when Bayes fails**
- ◆ When frequentists are stubborn and wrong
- ◆ Bayes factors
- ◆ Interim design analyses



Being Stubborn and Wrong

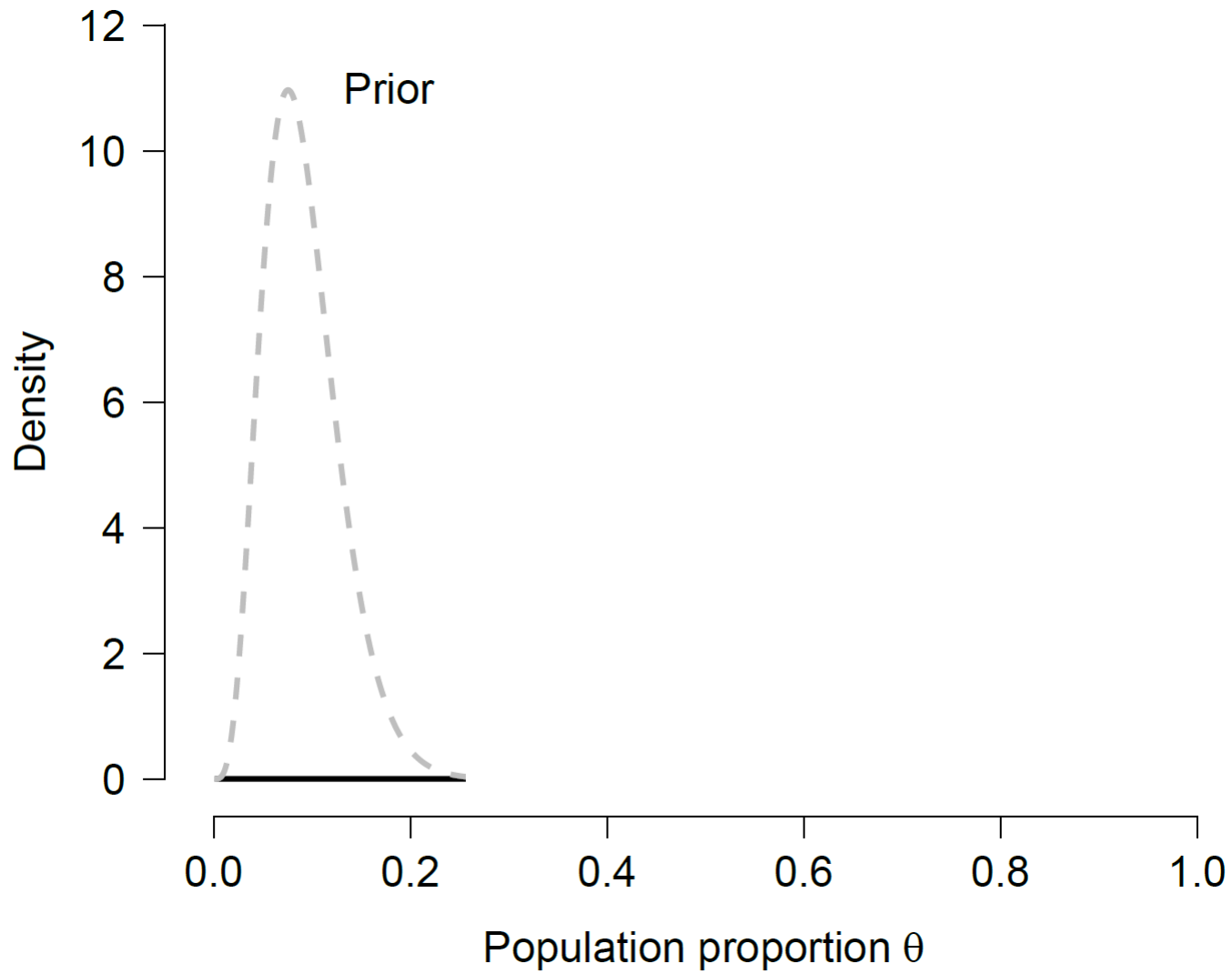
In a nutshell, a Bayesian will perform poorly if he/she is both misguided (with prior mean far from the true value of the parameter) and stubborn (placing a good deal of weight near the prior mean).

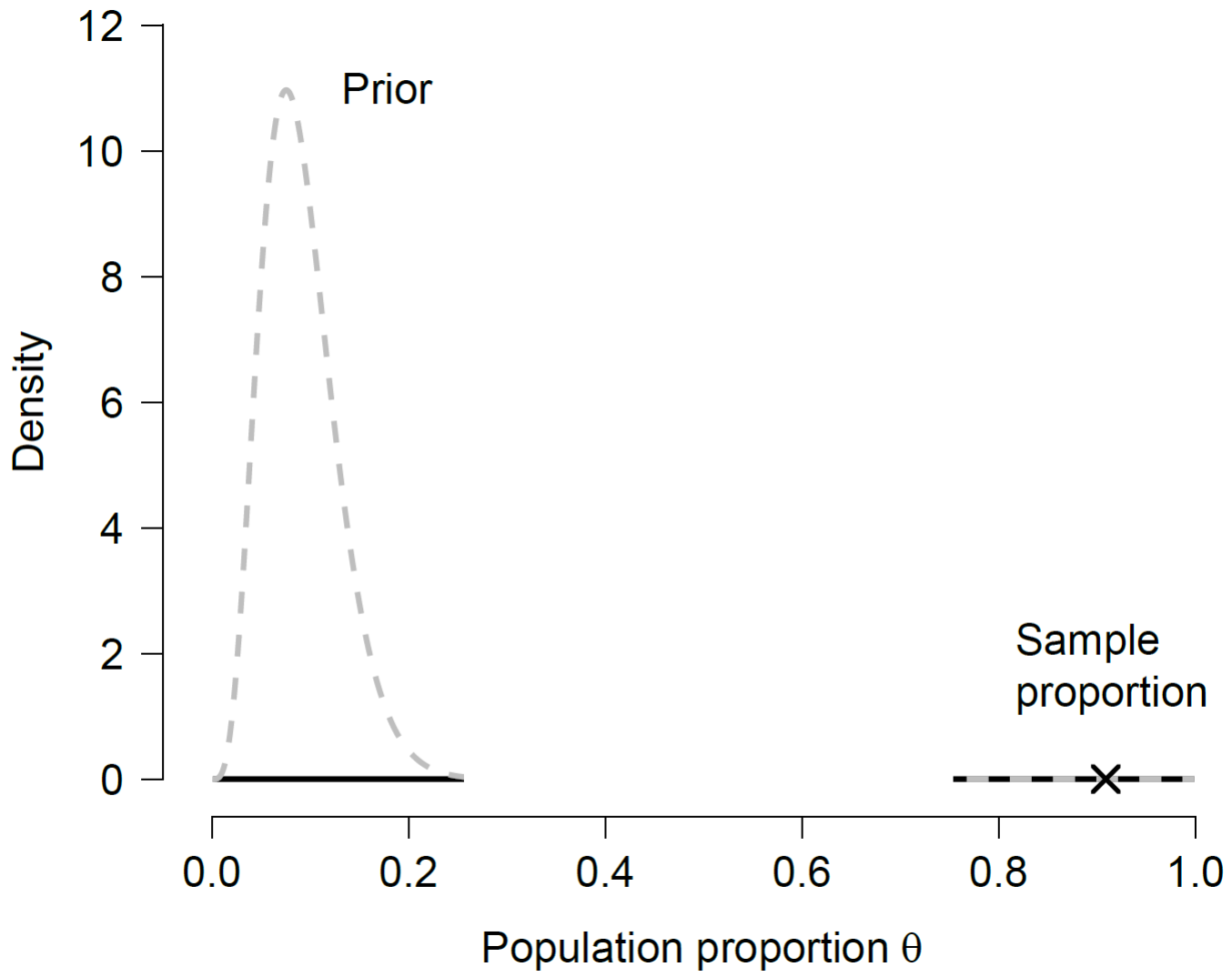
Samaniego, 2013

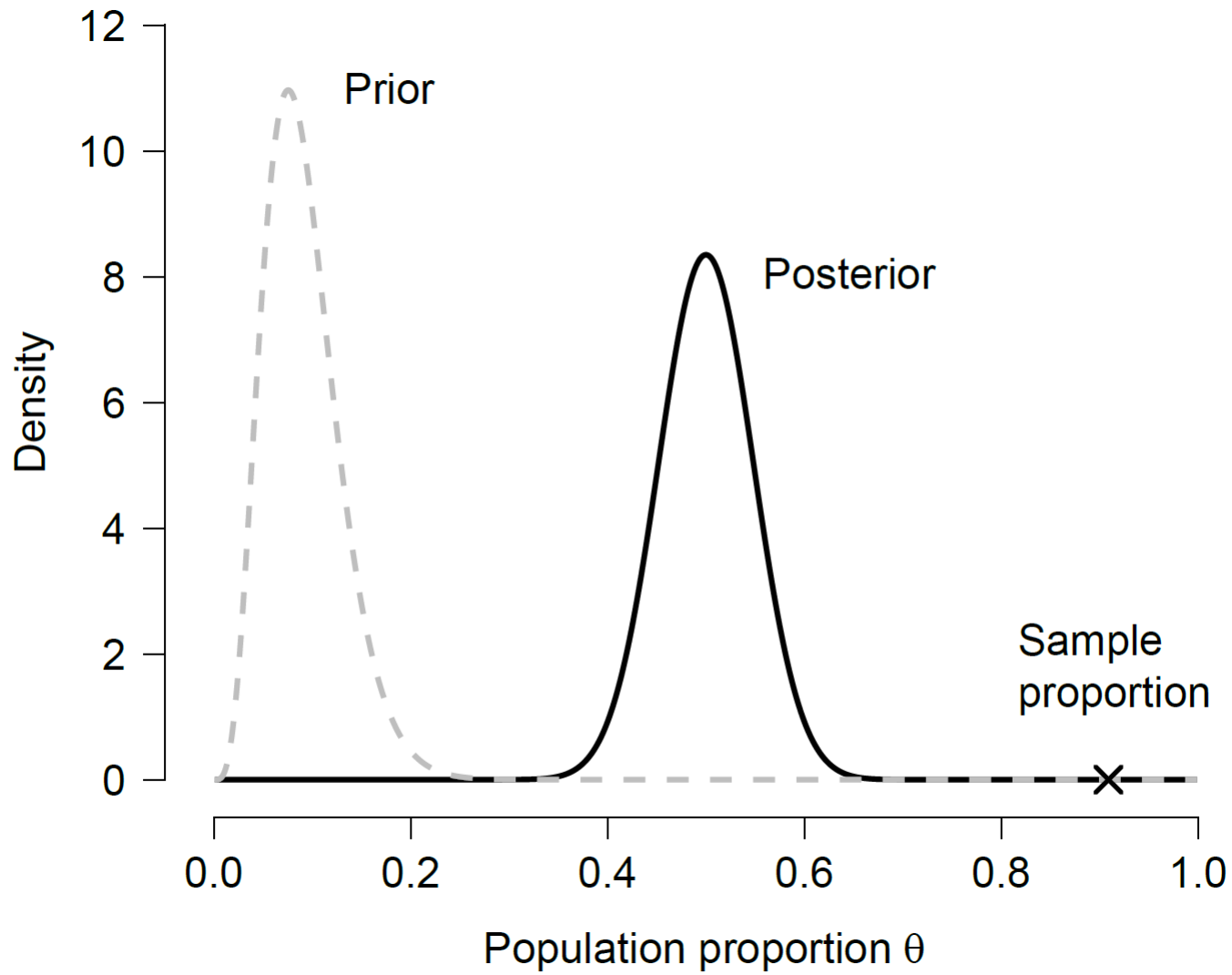


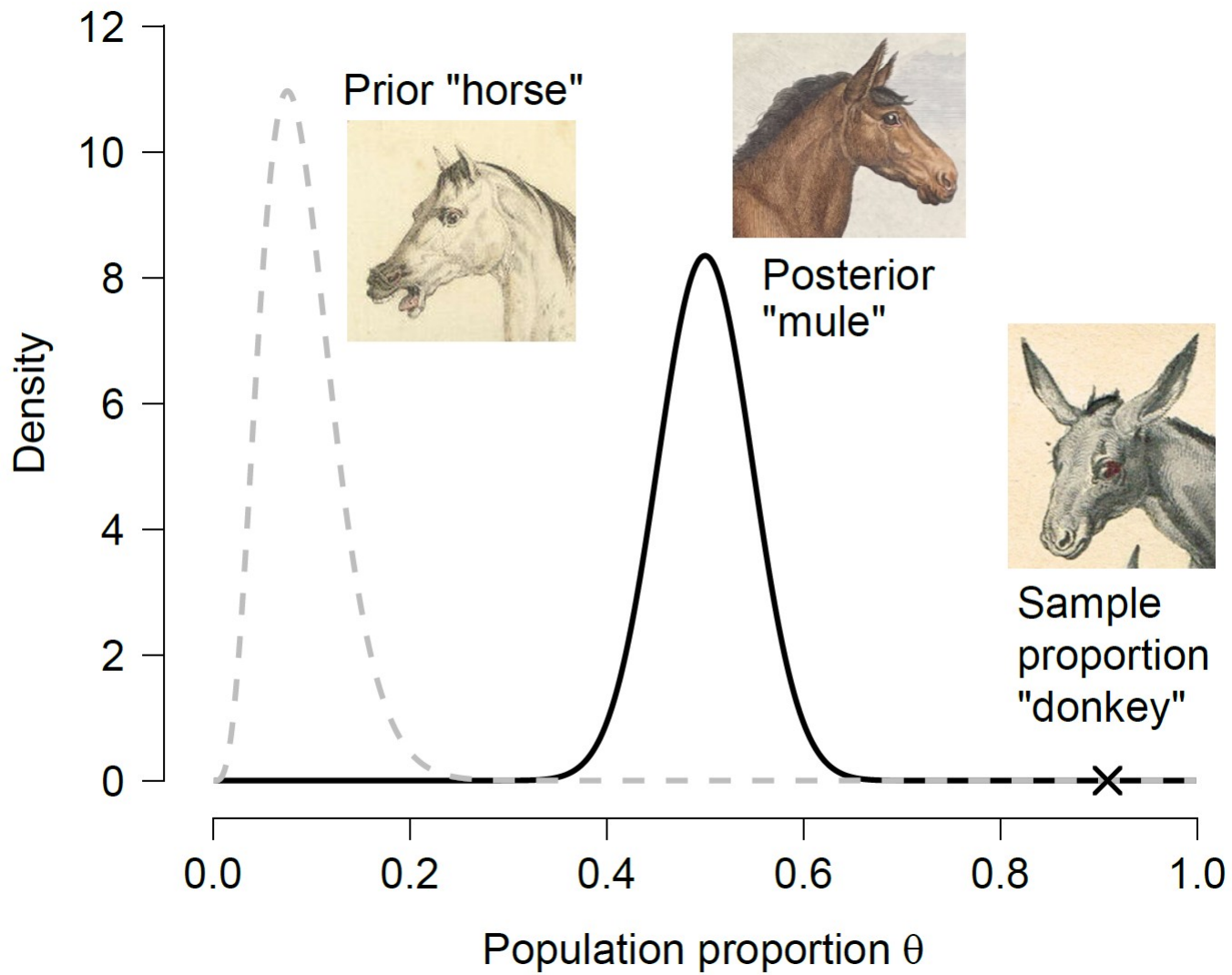
Definition of a Bayesian (Adjusted from Senn, 2007)

“One who, strongly expecting a horse and clearly viewing a donkey, confidently asserts having seen a mule.”





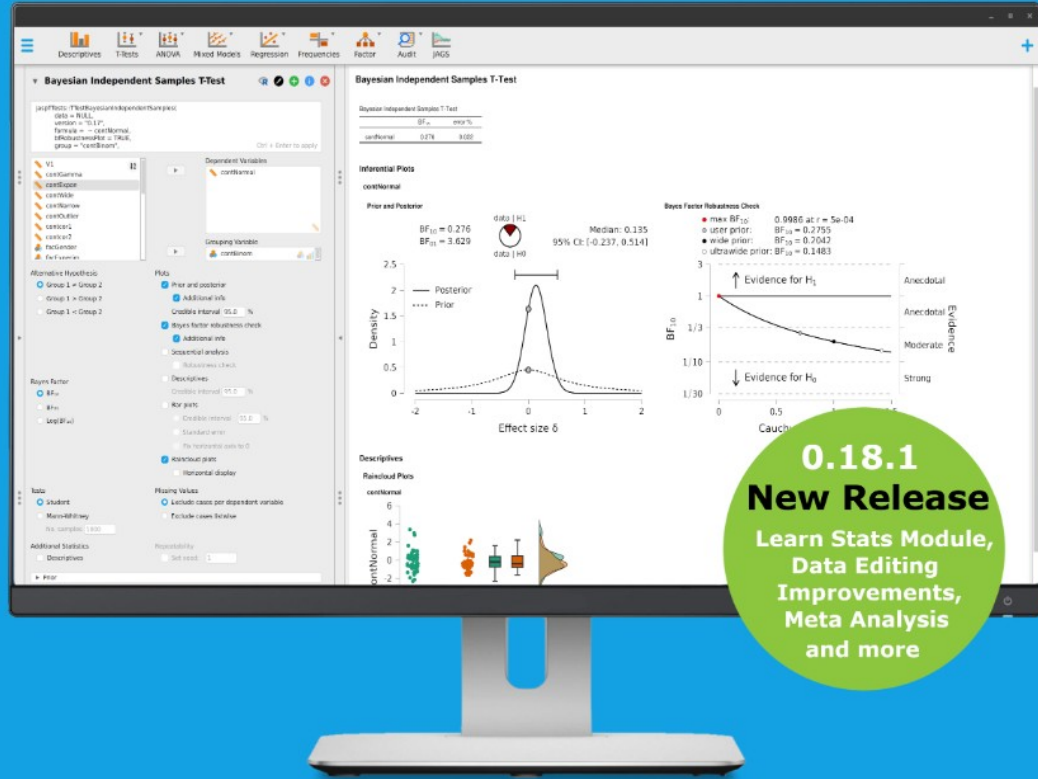




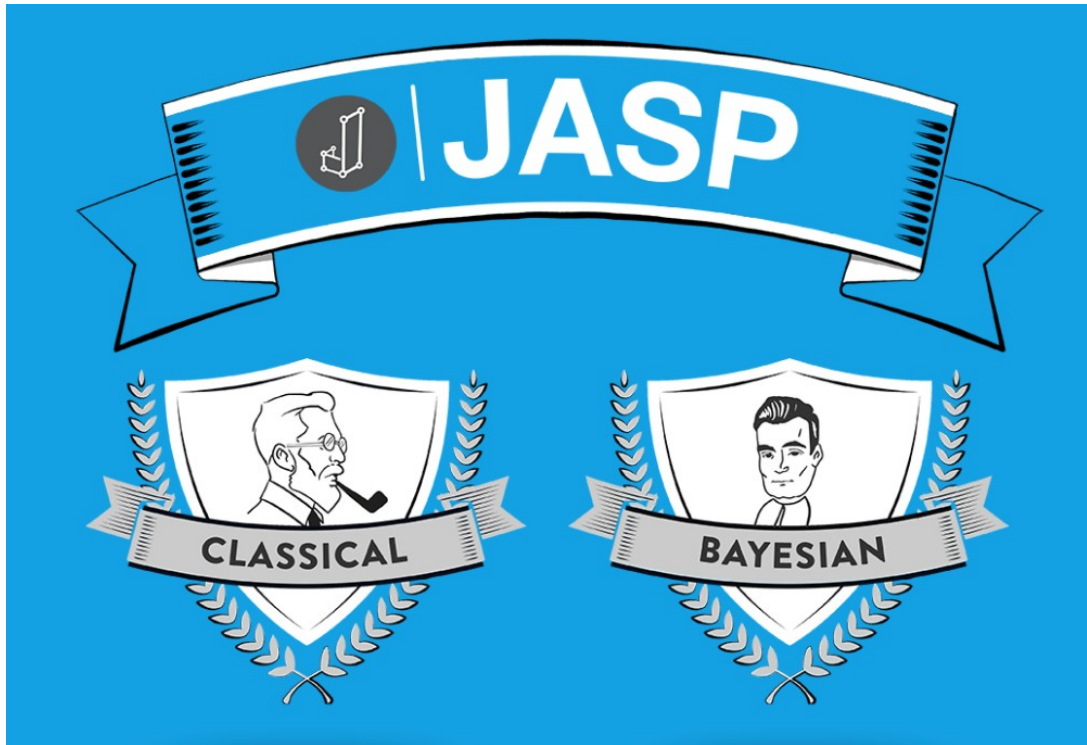


A Fresh Way to Do Statistics

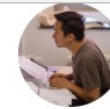
 Download JASP



More information at
jasp-stats.org



Eric-Jan Wagenmakers
CEO / Founder. Guides the development of JASP.
 @EJW [in](#)



Alexander Ly
CTO. Responsible for guiding JASP's scientific and technological strategy, and the choice of some Bayesian tests.
 @AL7



Bruno Boutin
Lead Software Developer. Responsible for the core development of JASP.
 @



Frans Meenhoff
Software Developer. Responsible for the core development of JASP.
 @



Akash Raj
Software Developer. Responsible for the implementation of UI elements, implemented the Summary Stats module.
 @



Quentin Gronau
Analyst. Responsible for the t-tests and the binomial test. Implemented the figures for the Bayesian analysis.
 @



Alexandra Sarafoglou
Analyst. Contributing to the multimonial analysis, the video tutorials, and the JASP workshop.
 @ [in](#)



Jan G. Voelkel
Software Developer. Responsible for improving the R analysis.
 @



Maarten Marsman
Analyst. Responsible for the Bayesian linear models (e.g., ANOVA and regression).
 @



Don van den Bergh
Analyst. Responsible for the frequentist and Bayesian reliability analysis, the machine learning module, and the network module. Also part of the workshop organization team.
 @



Johnny van Doorn
Analyst. Responsible for Bayesian nonparametric analysis and part of the workshop organization team.
 @



Dora Matzke
Analyst. Responsible for developing and maintaining the help, functionality, and the JASP documentation.
 @ [in](#)



Sacha Epskamp
Analyst. Responsible for factor analysis and the SEM module.
 @



Alexander Ezz
 The voice of many JASP video tutorials and other videos on our Youtube channel.
 @ [in](#)



Erik-Jan van Kesteren
Software developer. Responsible for adding plots, functions, and UI elements, and interfacing R and C++.
 @



Raouf Graman
Data scientist and code contributor. Responsible for improving code and developing new modules.
 @ [in](#)



Herbert Hoijtink
 Contributing to the Informative Hypotheses module.
 @



Joris Mulder
 Contributing to the Informative Hypotheses module.
 @



Yin Gu
 Contributing to the Informative Hypotheses module.
 @



Richard Morey
 Author and maintainer of the BayesFactor package.
 @



Tim Draws
Marketing and Communication Manager. Responsible for marketing strategy, website, blog, and the Youtube channel.
 @ [in](#)



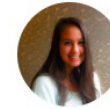
Koen Derks
 Contributing to the Machine Learning module, and the Bayesian Informative Hypothesis Testing module.
 @ [in](#)



Joris Goosen
Software developer. Responsible for the core development of JASP.
 @ [in](#)



Lotte Kehler
 Contributing to the blog, Youtube channel and manual of JASP.
 @ [in](#)



Tenth Annual JASP Workshop

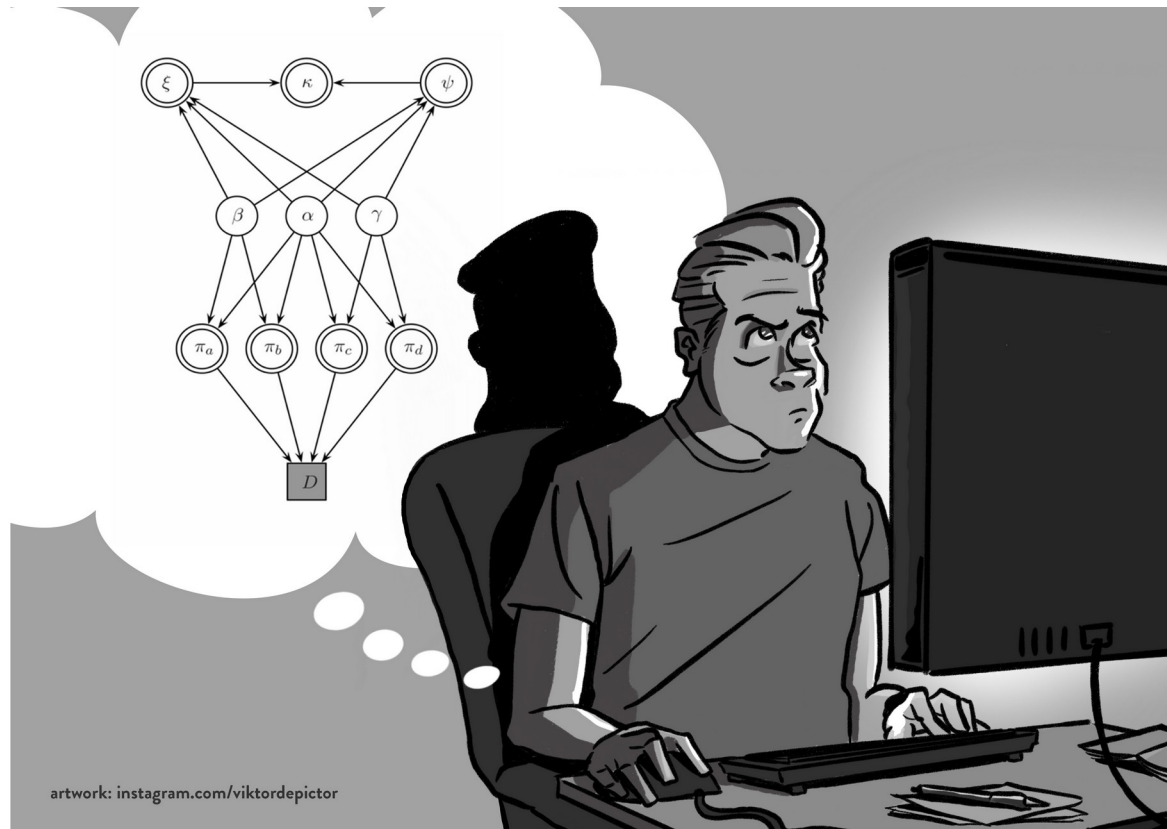
Theory and Practice of Bayesian Hypothesis Testing



June ?? & ??, 2024
University of Amsterdam

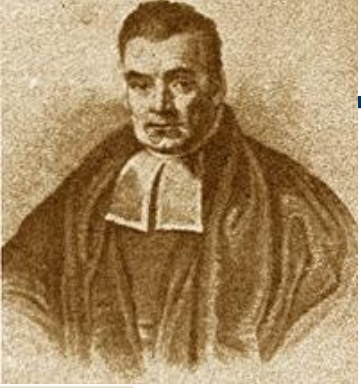
Twelfth Annual JAGS Workshop

Bayesian Modeling for Cognitive Science



June ??-??, 2024
University of Amsterdam





Overview

- ◆ Arguments pro Bayes
- ◆ Stubborn and wrong: when Bayes fails
- ◆ **When frequentists are stubborn and wrong**
- ◆ Bayes factors
- ◆ Interim design analyses



Frequentist Planning

- ◆ Assume a single population effect size δ under the alternative hypothesis H_1 ;
- ◆ Determine the sample size that gives a reasonable chance of correctly rejecting H_0 .



Frequentist Planning

- ◆ Assume a single population effect size δ under the alternative hypothesis H_1 ;
- ◆ Determine the sample size that gives a reasonable chance of correctly rejecting H_0 .
- ◆ But how should δ be chosen?



The Smallest Effect Size of Interest?

- ◆ Whose interest?



The Smallest Effect Size of Interest?

- ◆ Whose interest?
- ◆ What if we want to establish a theoretical causal connection, so any $\delta > 0$ suffices?
[Higg's boson, ESP]



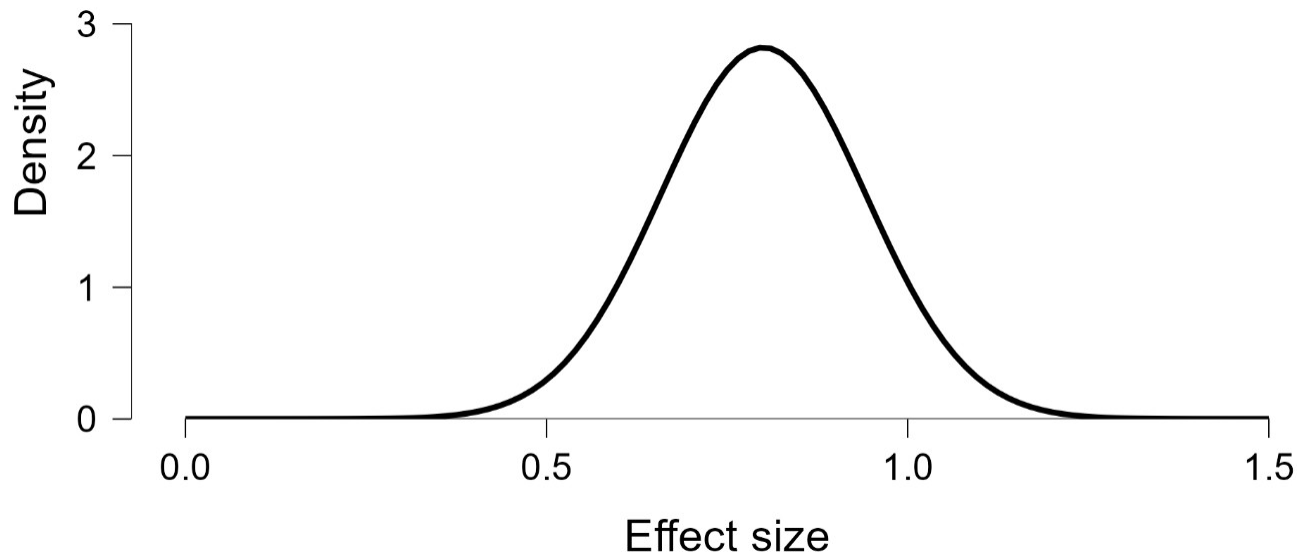
The Smallest Effect Size of Interest?

- ◆ Whose interest?
- ◆ What if we want to establish a theoretical causal connection, so any $\delta > 0$ suffices? [Higg's boson, ESP]
- ◆ Does method X affect the biological mechanism at all? [Does whiskey cure snake bite? If so we could enhance the dose for a better effect]



What if the SESOI is *Implausible*?

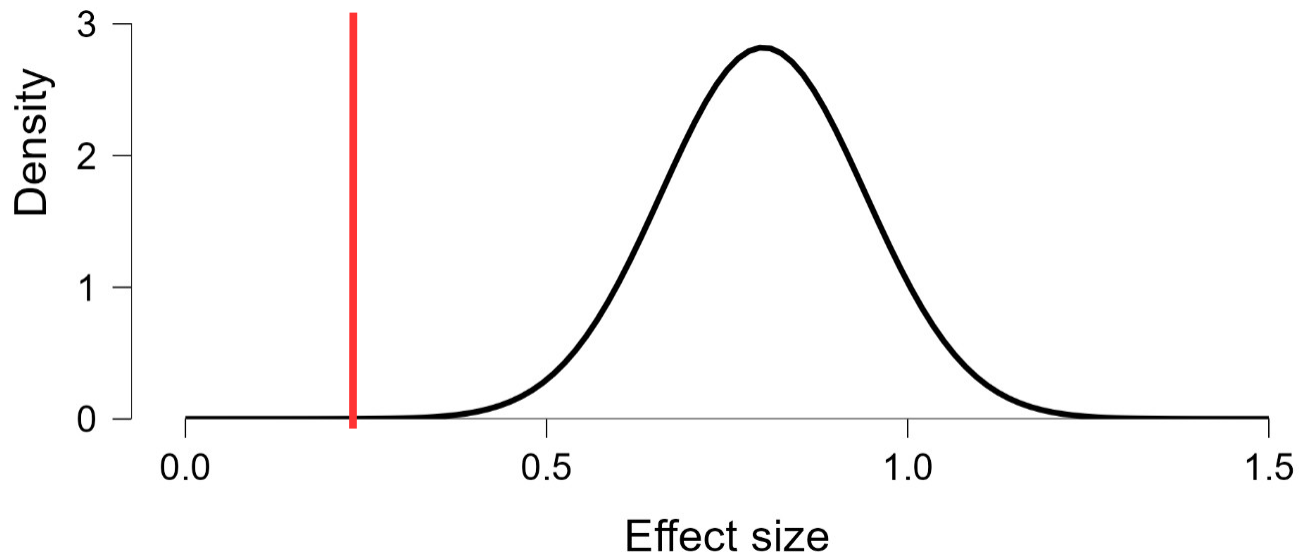
- ◆ SESOI implausibly small: experiment will likely be wasteful.





What if the SESOI is *Implausible*?

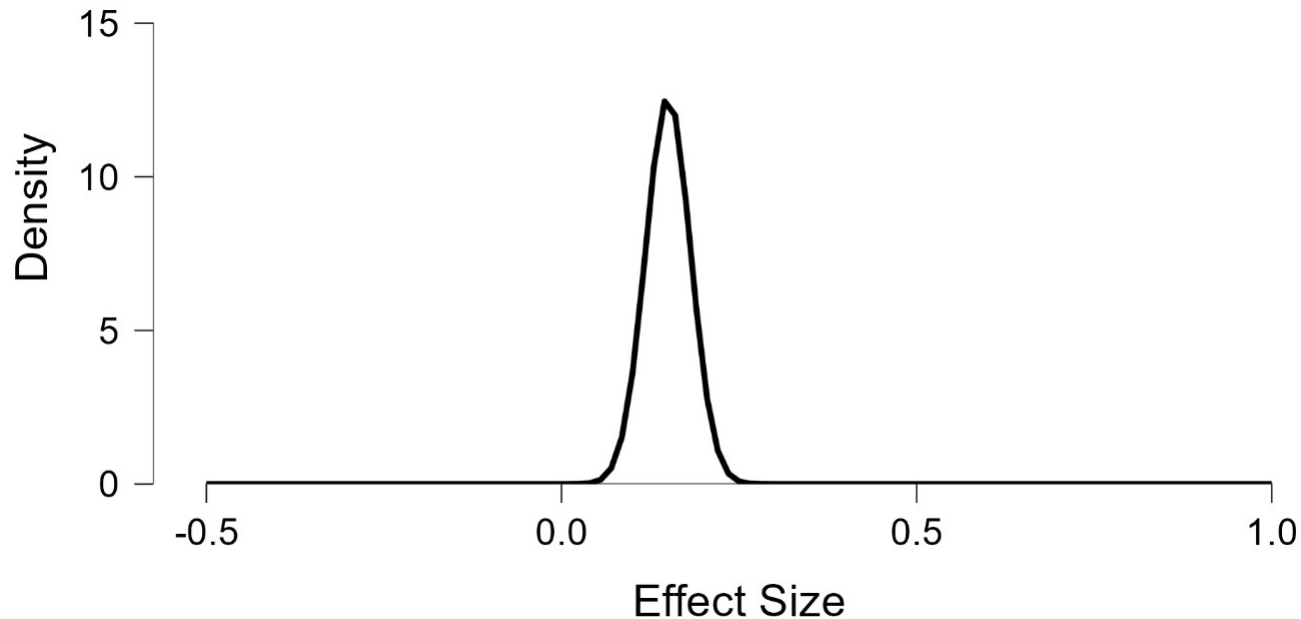
- ◆ SESOI implausibly small: experiment will likely be wasteful.





What if the SESOI is *Implausible*?

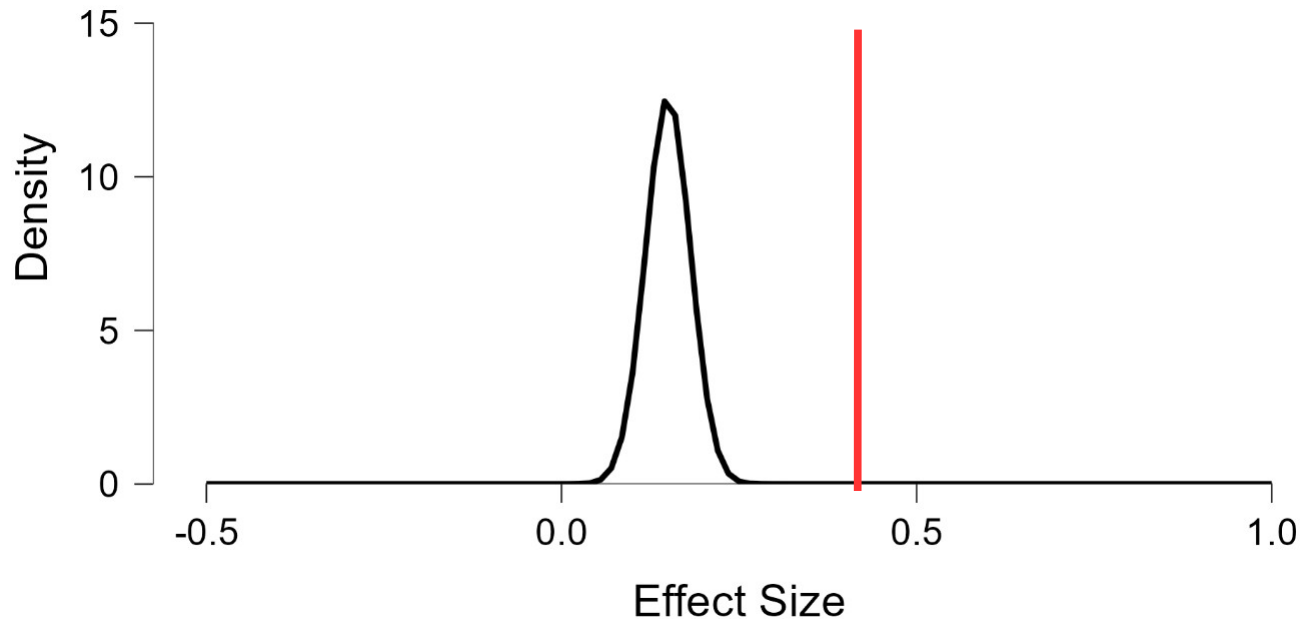
- ◆ SESOI implausibly large: experiment will run large risk of being nondiagnostic.





What if the SESOI is *Implausible*?

- ◆ SESOI implausibly large: experiment will run large risk of being nondiagnostic.





Take Home Message, I

- ◆ When planning a study, even frequentists must confront the issue of what effect sizes are *plausible*.
- ◆ The frequentist selects one value of δ for planning, and decides on a sample size.



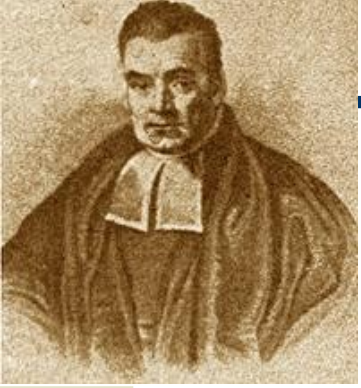
Take Home Message, I

- ◆ When planning a study, even frequentists must confront the issue of what effect sizes are *plausible*.
- ◆ The frequentist selects one value of δ for planning, and decides on a sample size.
- ◆ But what if this value proves to be completely wrong?




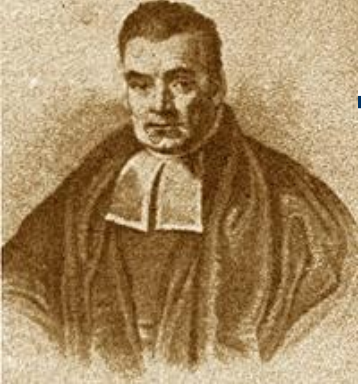
Take Home Message, II

- ◆ The frequentist now finds themselves in Senn's donkey scenario, but without the ability to learn *at all*.
- ◆ The single value of δ cannot be updated in-between; the experiment cannot be redesigned on the fly.
- ◆ There is no recovery from this, except to start all over. The frequentist is simply screwed.



Overview

- ◆ Arguments pro Bayes
 - ◆ Stubborn and wrong: when Bayes fails
 - ◆ When frequentists are stubborn and wrong
 - ◆ **Bayes factors**
 - ◆ Interim design analyses
- 
- A decorative vertical bar on the left side of the slide, consisting of a series of horizontal lines of varying lengths, creating a textured, striped effect.



Bayes Factors: Data-Driven Change in Beliefs

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

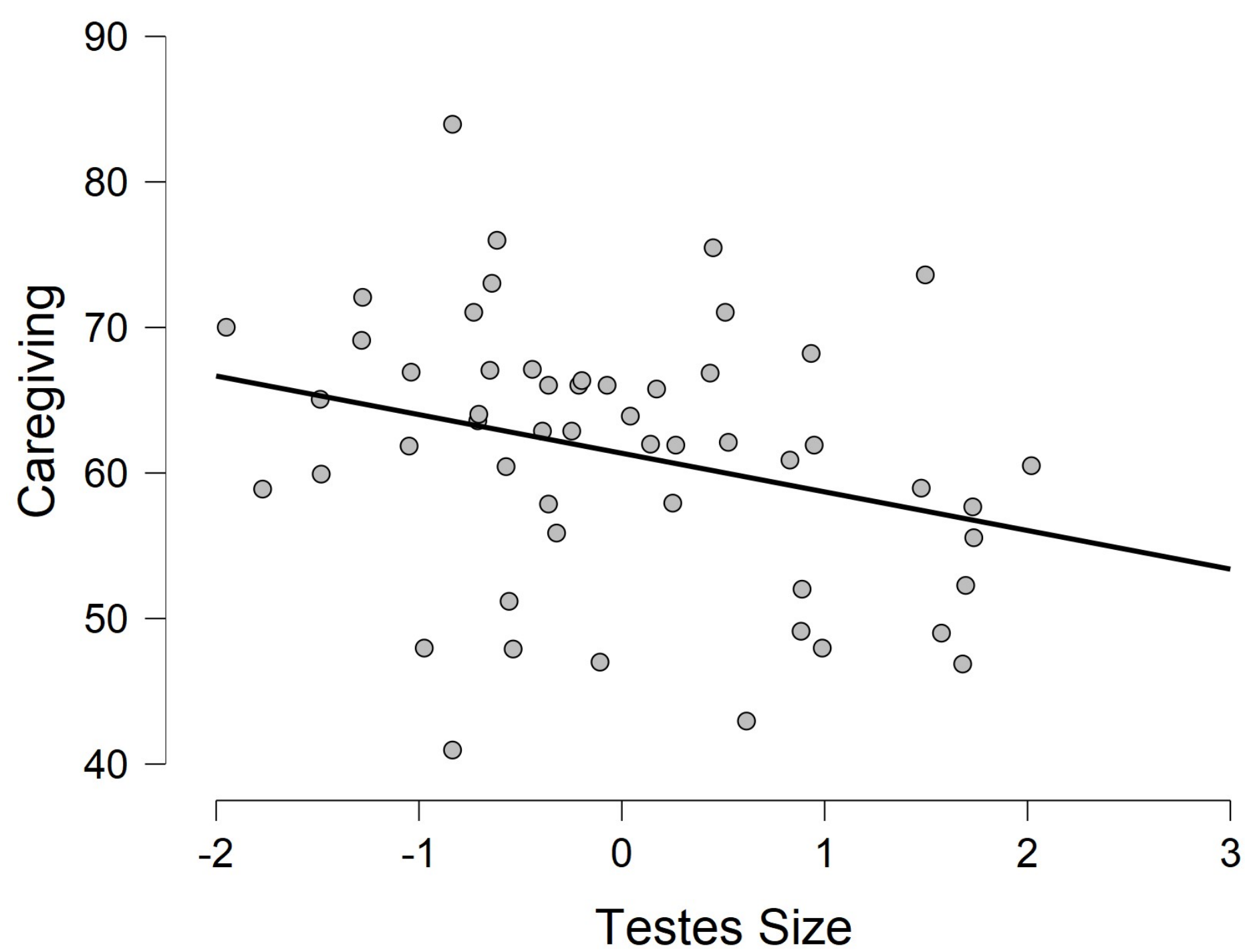
Testicular volume is inversely correlated with nurturing-related brain activity in human fathers

Jennifer S. Mascaro^{a,b,c}, Patrick D. Hackett^a, and James K. Rilling^{a,b,c,d,1}

15746–15751 | PNAS | September 24, 2013 | vol. 110 | no. 39

Results

Reproductive Biology and Parenting Behavior. Although testes volume was not related to body mass, there was a significant linear correlation between testes volume and height [$r(53) = 0.27, P < 0.05$]. Therefore, residual testes volume, controlling for height, was used in subsequent analyses. Residual testes volume was negatively related to paternal caregiving [$r(52) = -0.29, P < 0.05$]



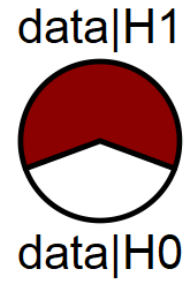
Discussion

Collectively, these data provide the most direct support to date that the biology of human males reflects a trade-off between mating and parenting effort. Fathers' testicular volume and testosterone levels were inversely related to parental investment

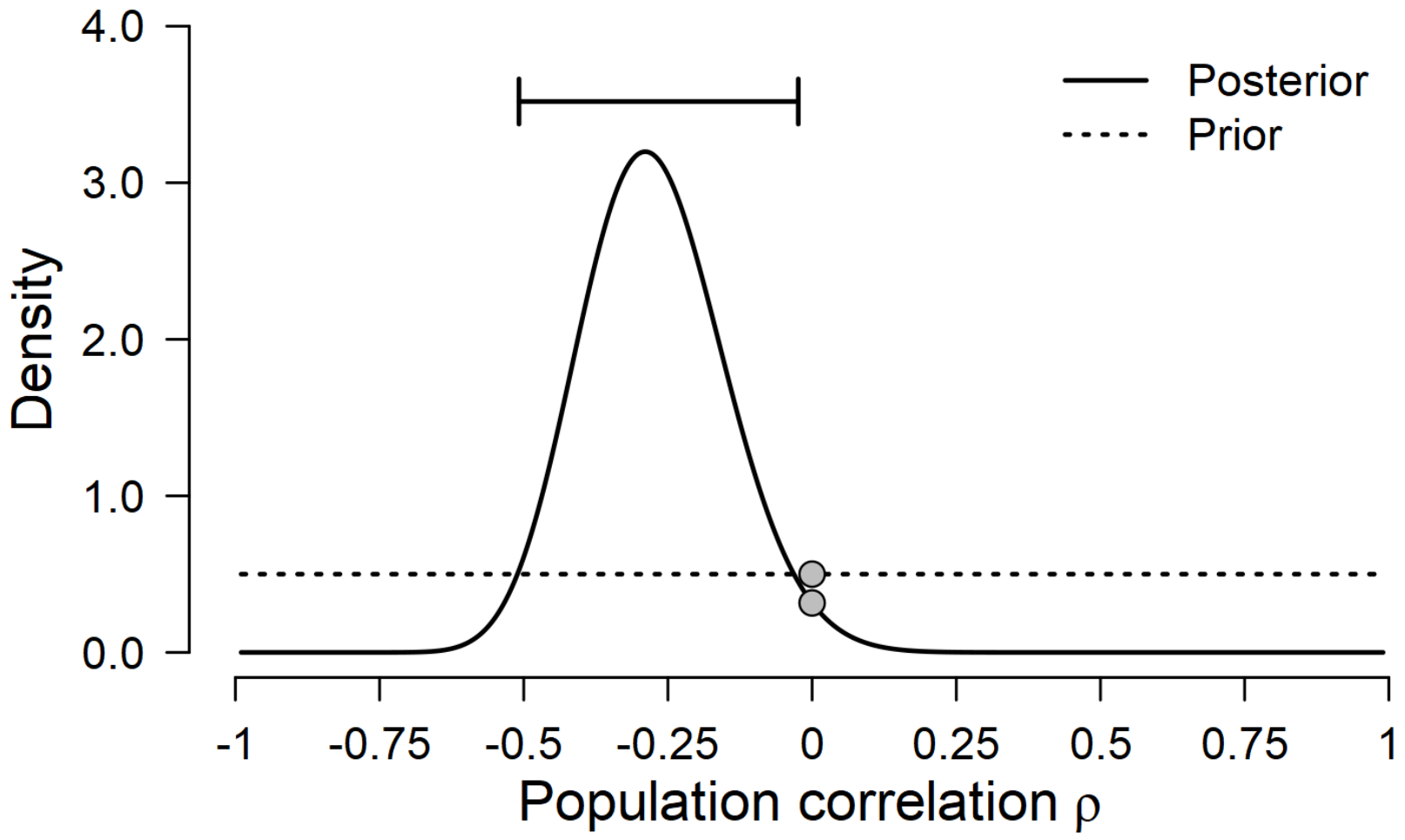


| **JASP**

$BF_{10} = 1.582$
 $BF_{01} = 0.632$



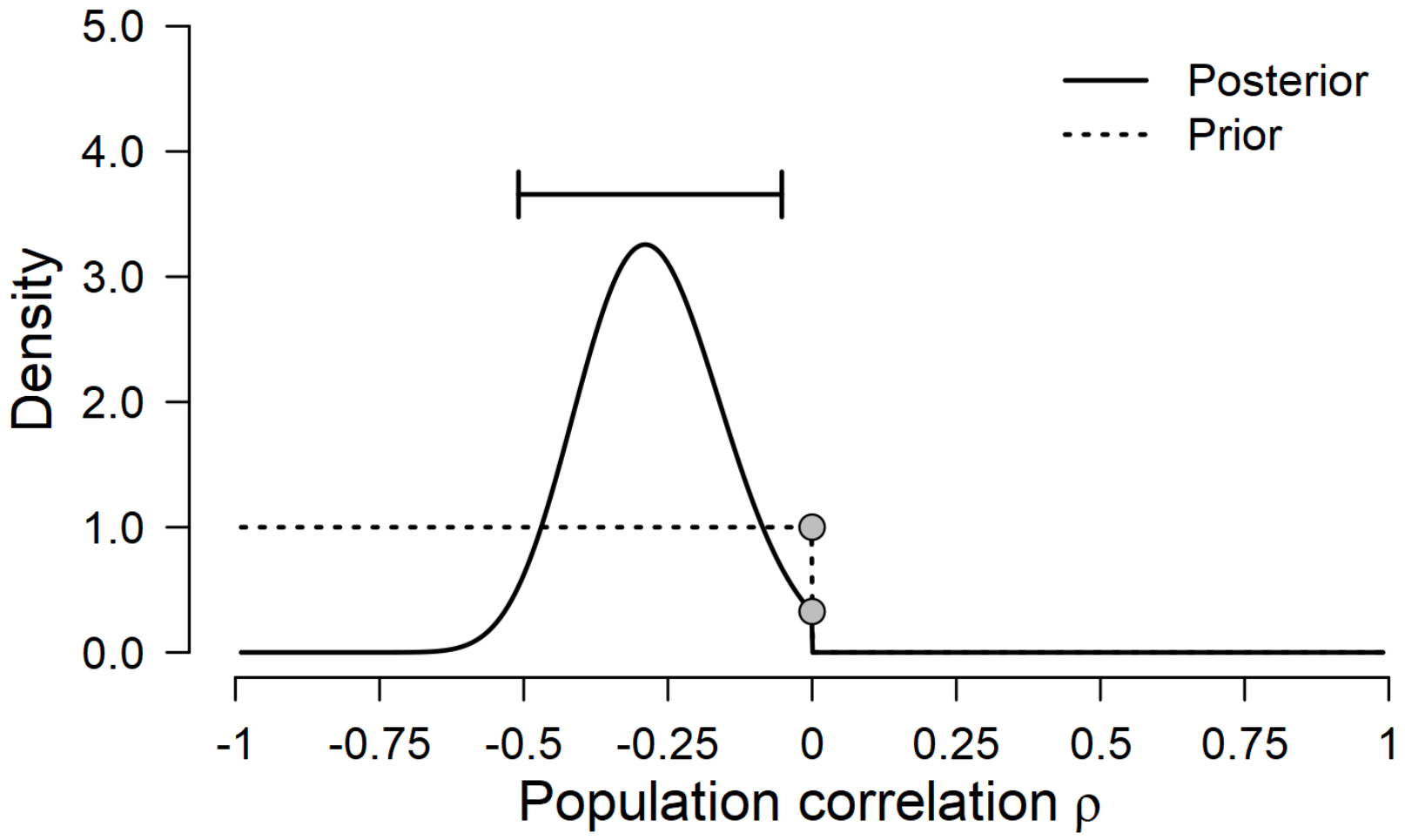
median = -0.278
95% CI: [-0.508, -0.024]



$BF_{-0} = 3.108$
 $BF_{0-} = 0.322$



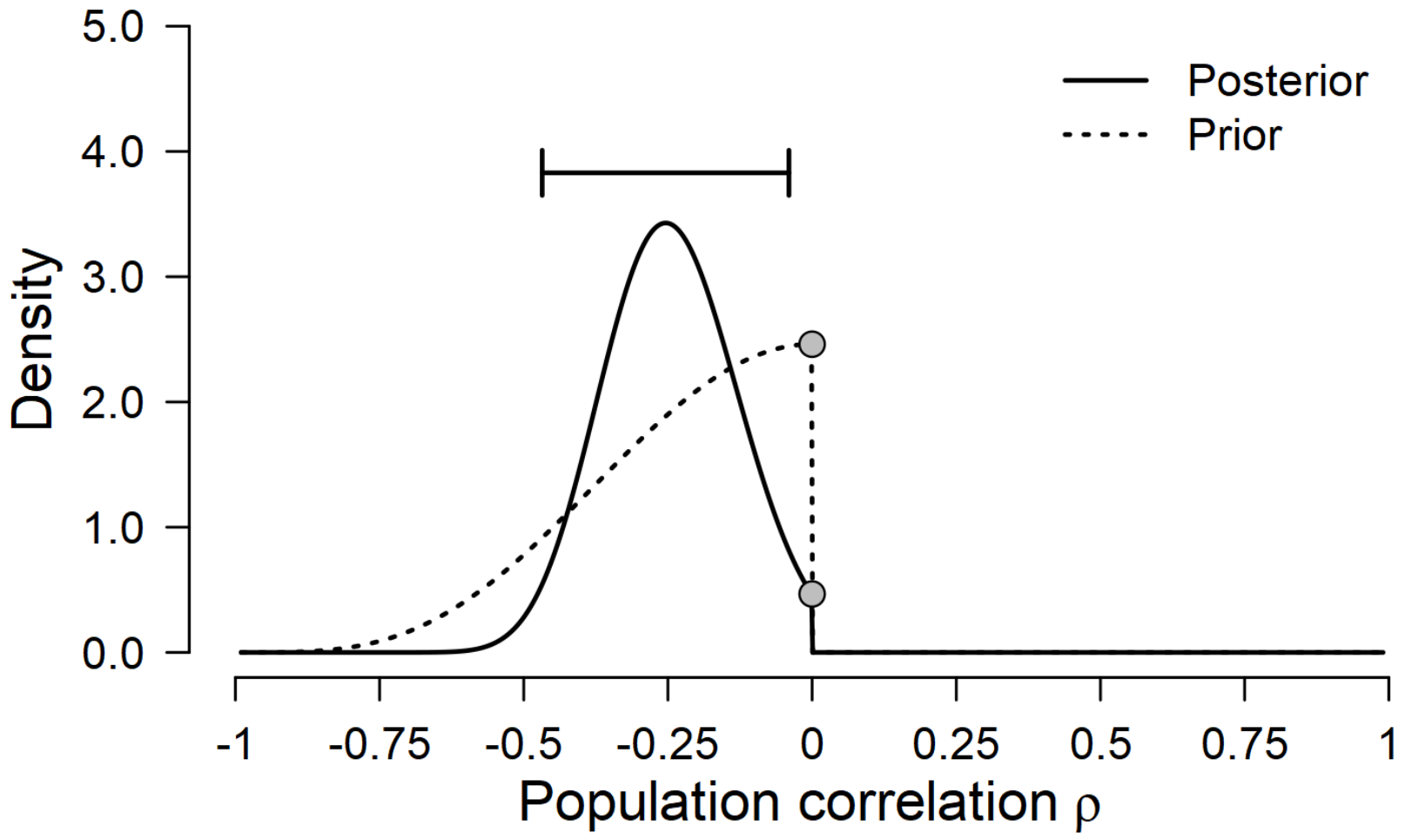
median = -0.281
95% CI: [-0.509, -0.053]



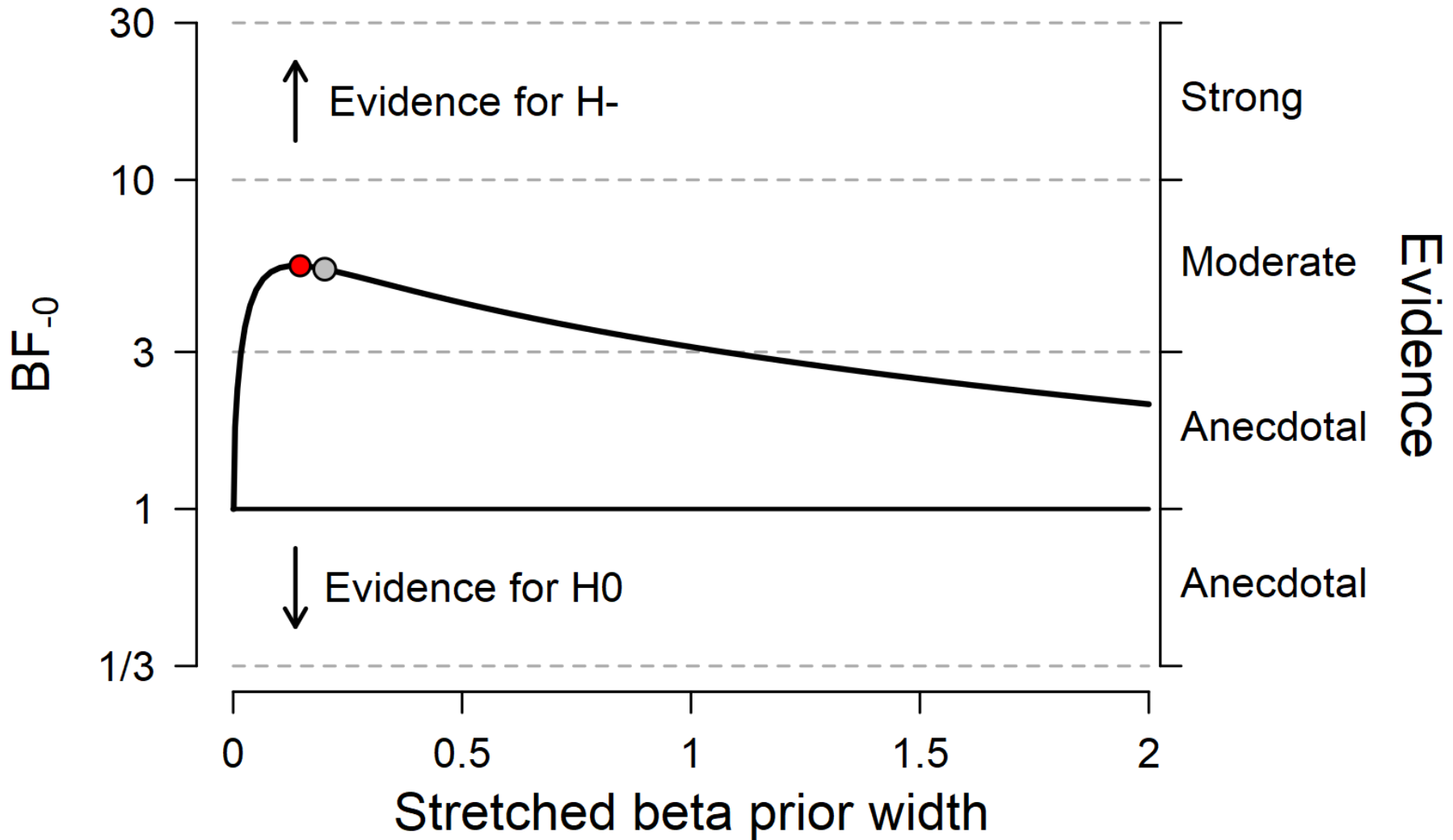
$BF_{-0} = 5.353$
 $BF_{0-} = 0.187$



median = -0.248
95% CI: [-0.468, -0.040]



- max BF_{-0} : 5.482 at $r = 0.1462$
- user prior: $BF_{-0} = 5.353$





Conclusion for PNAS Study: A Type B Error

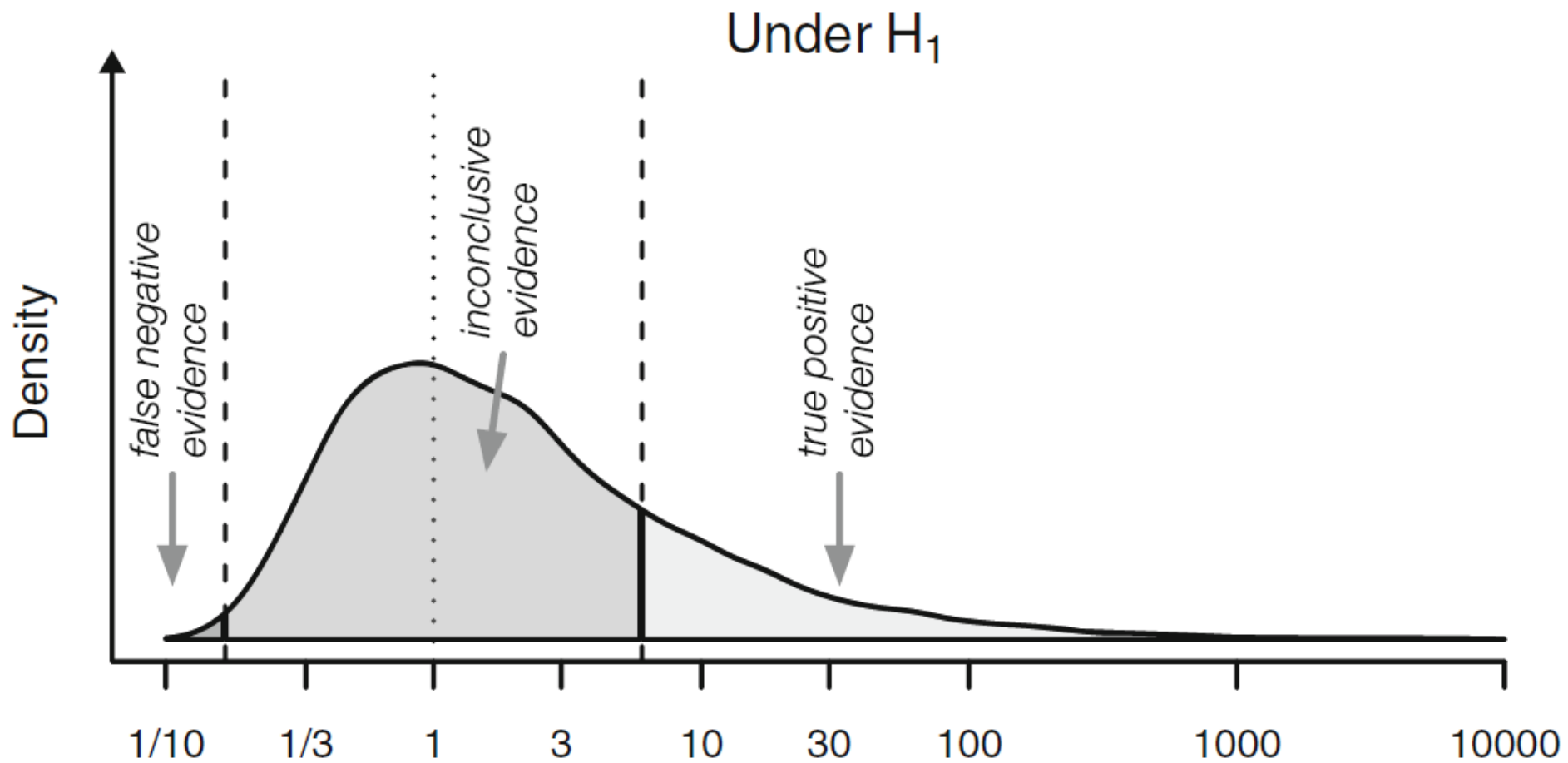
- ◆ The strength of evidence provided by the Bayes factor is weak-to-modest, and conflicts with the frequentist all-or-none decision to “reject the null hypothesis”.

Bayes Factor Design Analysis: Planning for Compelling Evidence

- ◆ We may design a study such that the probability of obtaining compelling evidence is relatively high.
 - Fix n , assess distribution on BFs
 - Fix BF, assess distribution on n

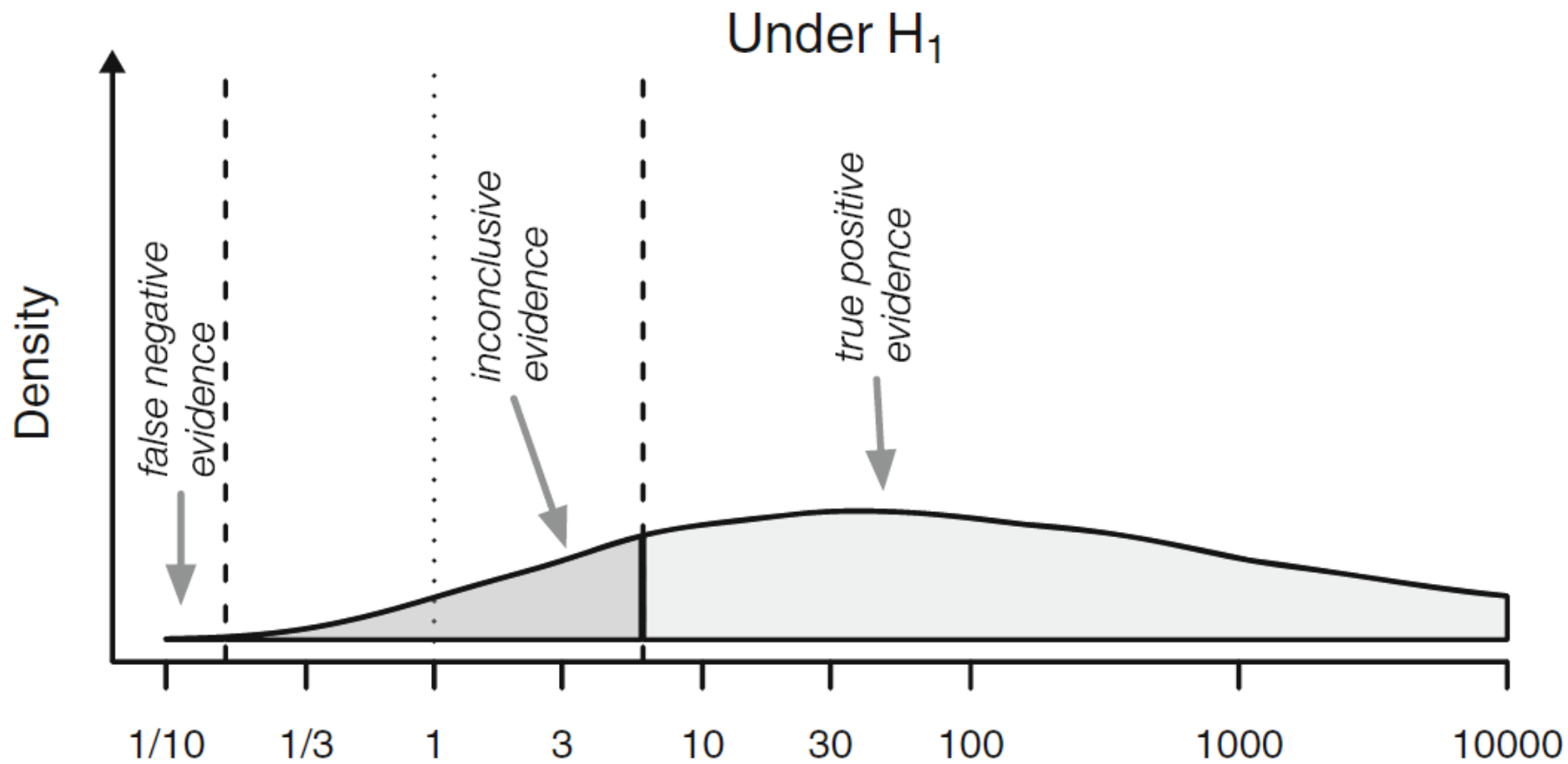
a

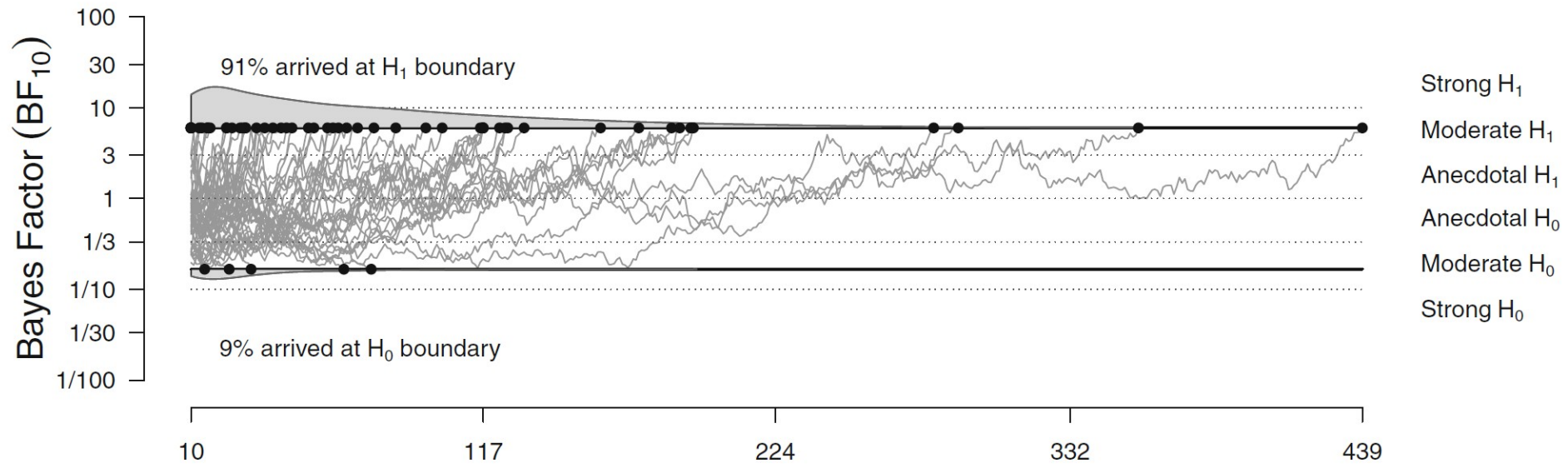
$$n = 20, \delta = 0.5$$



b

$$n = 100, \delta = 0.5$$





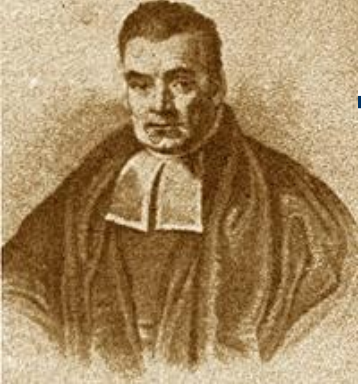
Bayes factor design analysis: Planning for compelling evidence

Felix D. Schönbrodt¹ · Eric-Jan Wagenmakers²

Behavior Research Methods (2019) 51:1042–1058
<https://doi.org/10.3758/s13428-018-01189-8>

A tutorial on Bayes Factor Design Analysis using an informed prior

Angelika M. Stefan¹ · Quentin F. Gronau¹ · Felix D. Schönbrodt² · Eric-Jan Wagenmakers¹

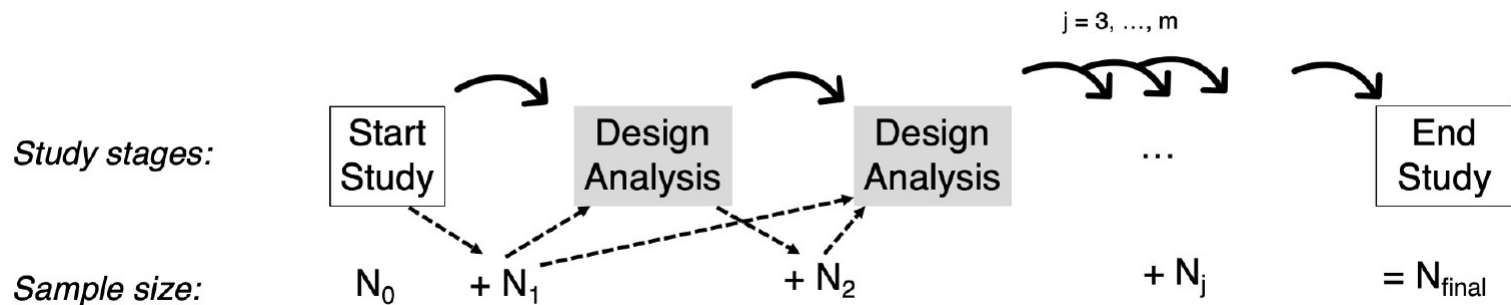


Overview

- ◆ Arguments pro Bayes
- ◆ Stubborn and wrong: when Bayes fails
- ◆ When frequentists are stubborn and wrong
- ◆ Bayes factors
- ◆ Interim design analyses

Key Insight: BFDA May Be Executed on the Fly

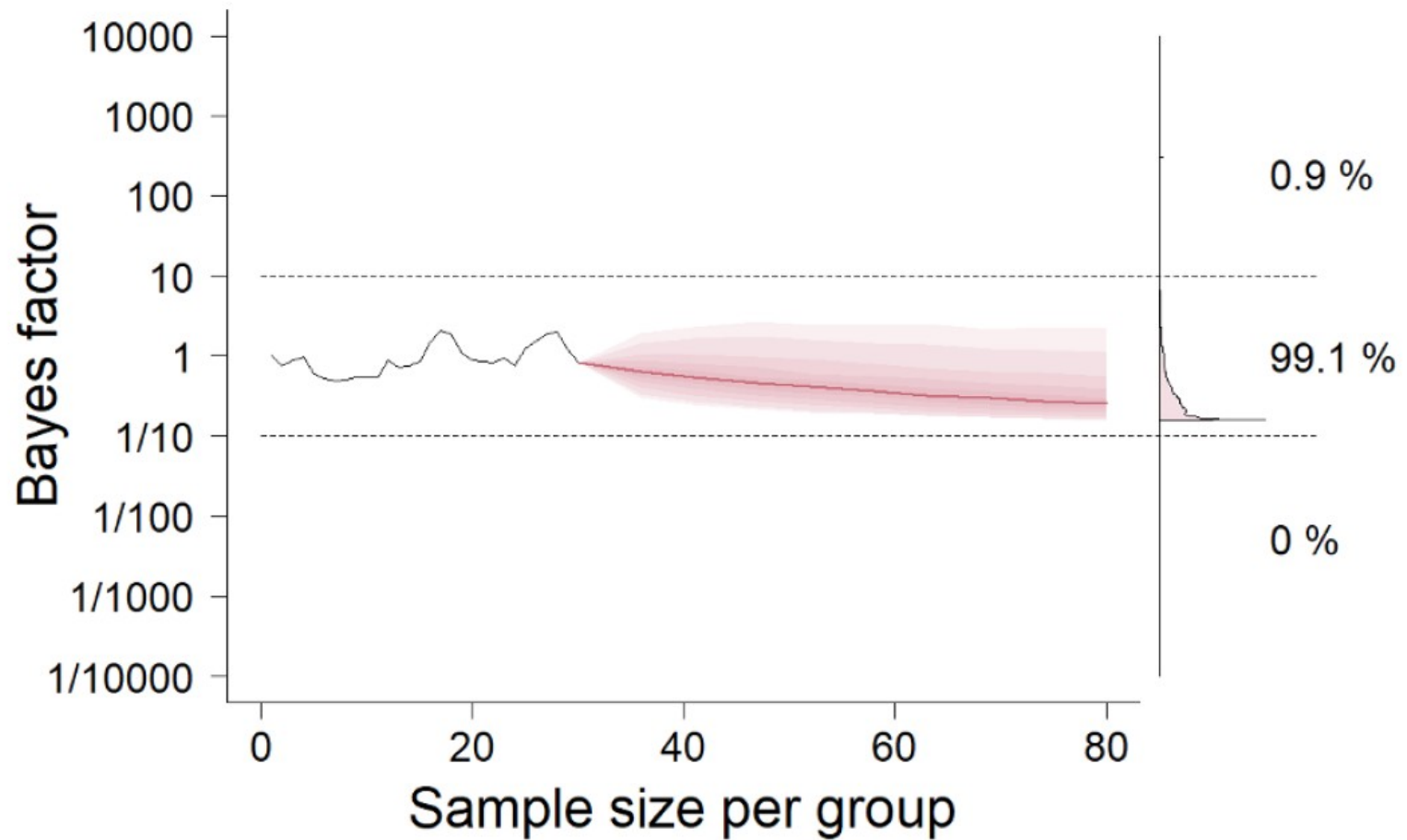
- ◆ As the data accumulate, we learn about the values of δ that are plausible.
- ◆ At any time we may conduct a new BFDA to quantify our updated expectations regarding evidence and sample size.



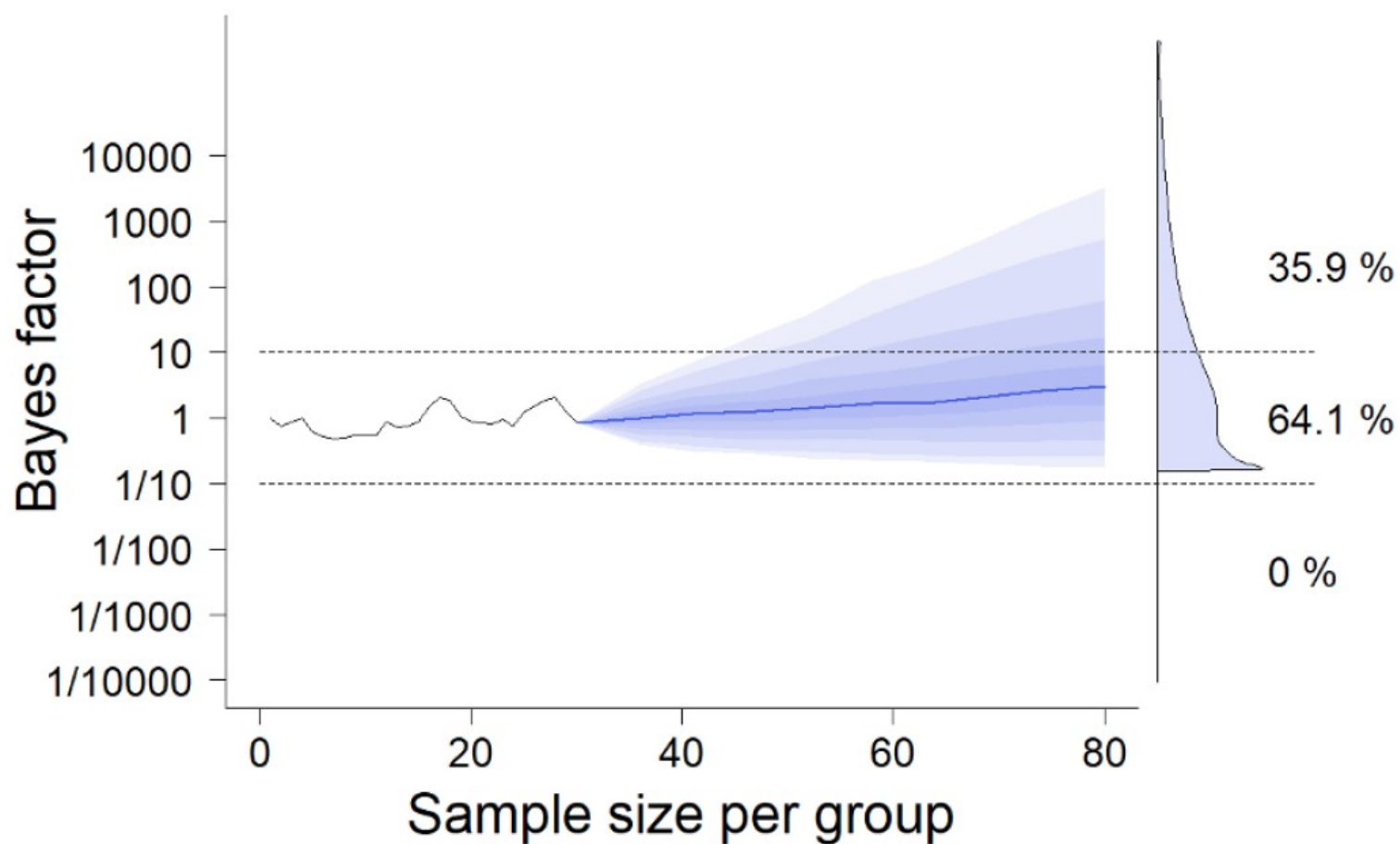
Interim Design Analysis Using Bayes Factor Forecasts

Angelika M. Stefan^{a,b}, Quentin F. Gronau^c, Eric-Jan Wagenmakers^a

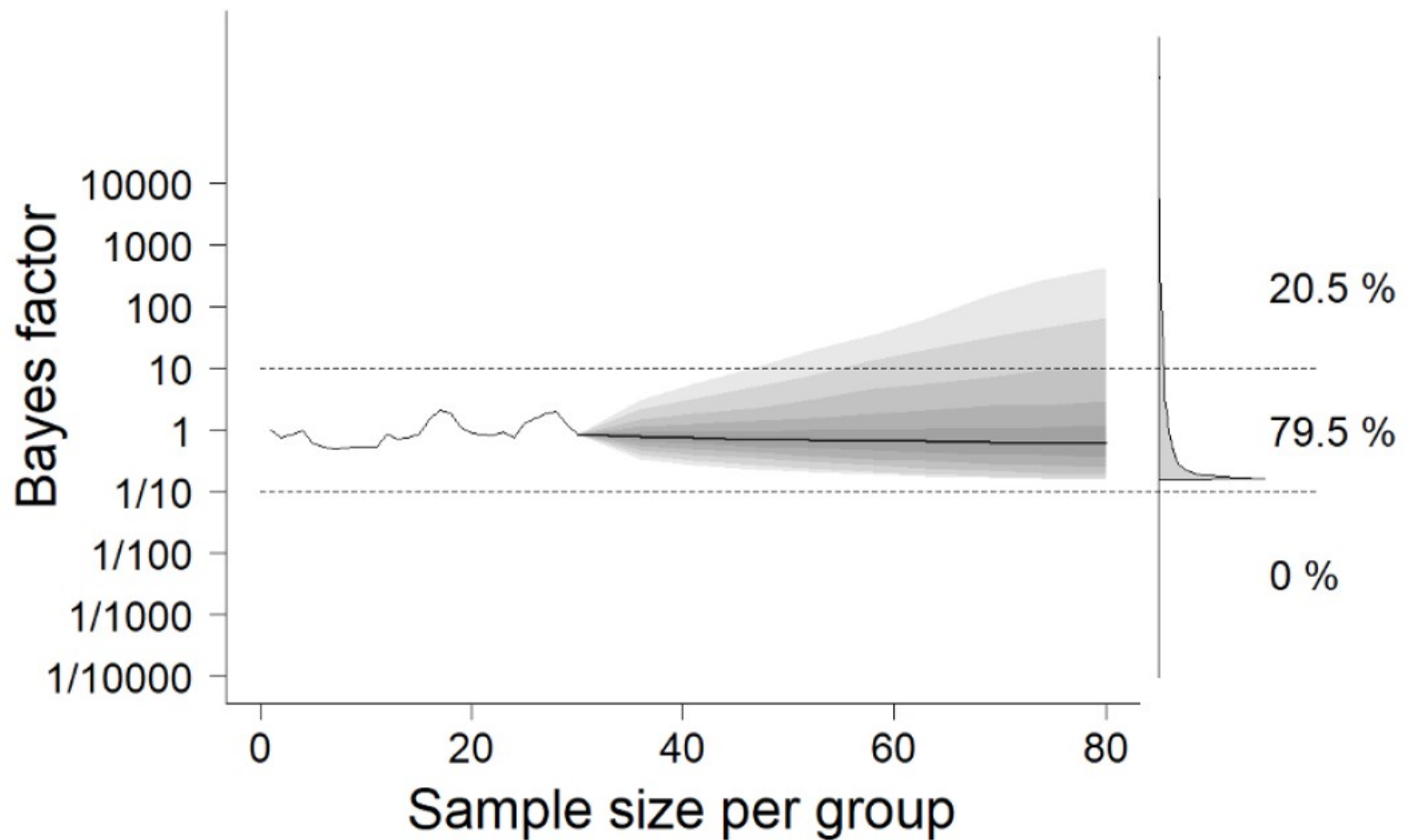
Bayes factor forecast under \mathcal{M}_0



Bayes factor forecast under \mathcal{M}_1



Model-averaged Bayes factor forecast



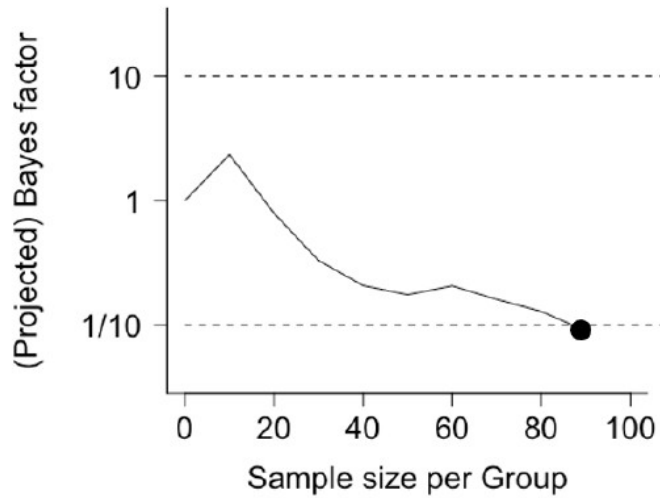


Reasons for Stopping

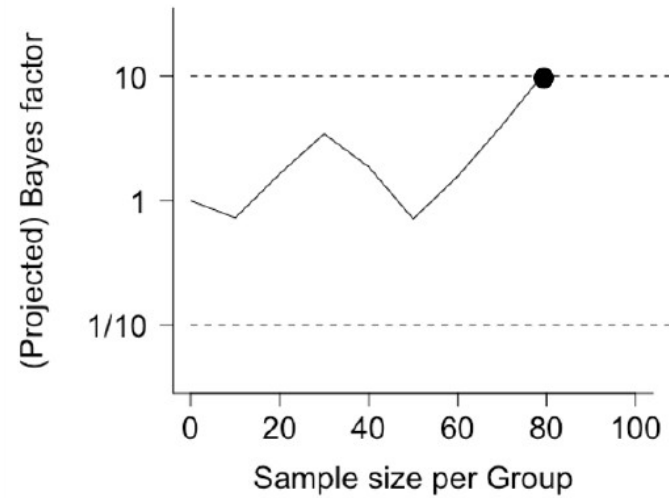


- ◆ Compelling evidence either way
- ◆ Resources depleted
- ◆ Futility

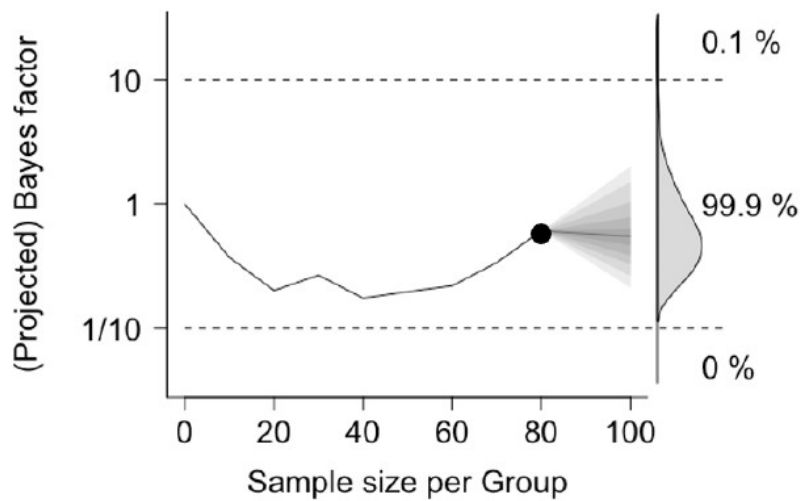
Strong evidence for \mathcal{M}_0



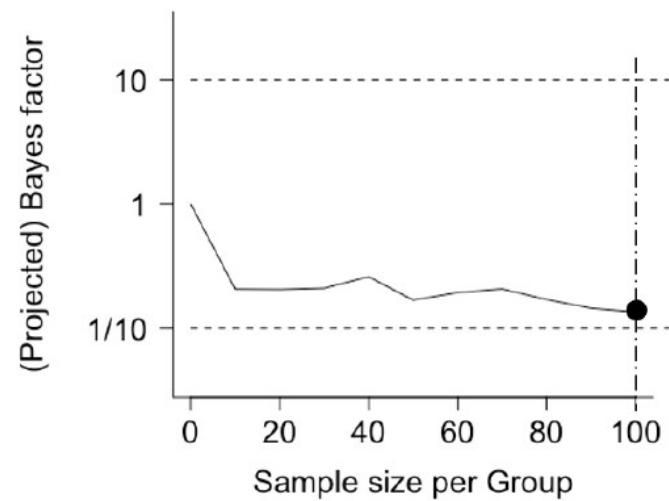
Strong evidence for \mathcal{M}_1

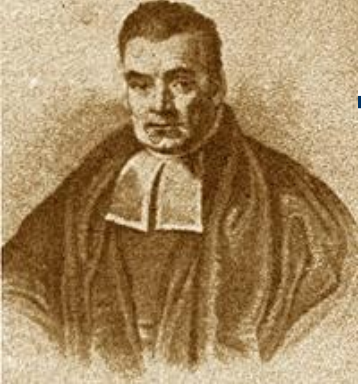


Futility



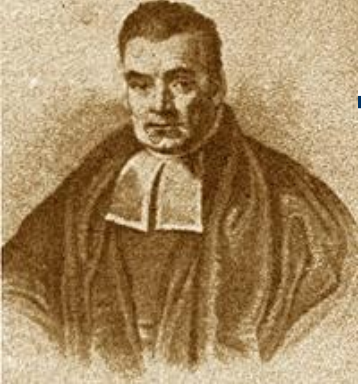
Maximum N





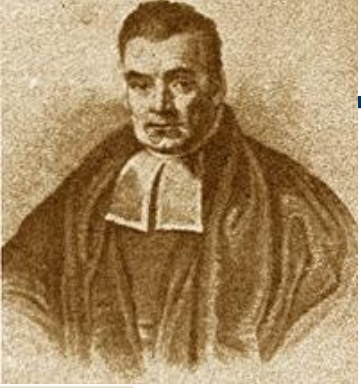
Conclusions I

- ◆ Bayesian inference is theoretically attractive, but also affords great *practical* advantages.
- ◆ I believe it is counterproductive for regulators to ignore Bayesian analyses. You may ask “what about Type I error control?”, but instead ask “what about the evidence?”



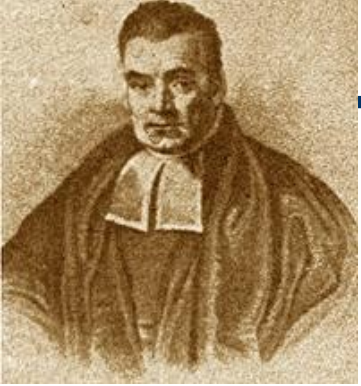
Conclusions II

- ◆ Key questions:
 - “*in light of the data, what is the probability that the treatment is effective?*”
 - “*how much do the data enhance the credibility of H_1 versus H_0 ?*”
- ◆ These fundamental questions can only be answered by a Bayesian analysis.



Conclusions III

- ◆ Another key question is “*in light of the data, should we allow this drug on the market ?*”
- ◆ This is *also* fundamentally a Bayesian question! Rational decision making requires that we bring together prior knowledge, data, and *utilities*.



Conclusions IV

- ◆ Instead of focusing on Type I errors, regulators ought to start worrying about:
 - *Type B errors*, where a reasonable Bayesian analysis contradicts the frequentist analysis;
 - *Type F errors*, where frequentist analyses are misused to answer Bayesian questions.



Thanks for your Attention!