My current views over the Bayes factor

Jorge N. Tendeiro
August 20, 2019

University of Groningen

Today's talk

I will present results from three papers, all revolving around the Bayes factor:

- Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*. https://doi.org/10.1037/met0000221. Preprint here: https://osf.io/t5xfd.

- Kiers, H. A. L. & Tendeiro, J. N. (2019). With Bayesian estimation one can get all that Bayes factors offer, and more. Submitted. Preprint here: https://psyarxiv.com/zbpmy

- Tendeiro, J. N., Kiers, H. A. L., & van Ravenzwaaij, D. (2019). Tentative title: A mathematical proof for optional stopping using NHBT. Close to submit (no preprint yet!).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Motivation

"The field of psychology is experiencing a crisis of confidence, as many researchers believe published results are not as well supported as claimed."[1]

Q: Why?

A: Among several other reasons (QRPs[2,3]), due to overreliance on NHST and p-values.[4,5,6,7]

[1] Rouder (2014).
[2] John, Loewenstein, and Prelec (2012).
[3] Simmons, Nelson, and Simonsohn (2011).
[4] Edwards, Lindman, and Savage (1963).
[5] Cohen (1994).
[6] Nickerson (2000).
[7] Wagenmakers (2007).

Bayes factors are being increasingly advocated as a better alternative to NHST.[1,2,3,4,5]

We felt we did not know enough about Bayes factors (peculiarities, pitfalls, problems).

We surveyed the literature. Here we summarize what we found.

---

[1] Jeffreys (1961).
[2] Wagenmakers et al. (2010).
[3] Vampaemel (2010).
[4] Masson (2011).
[5] Dienes (2014).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Definition

Definition

The Bayes factor[1,2] quantifies the change in prior odds to posterior odds due to the data observed.

- Two models to compare, for instance $\mathcal{M}_0 : \theta = 0$ vs $\mathcal{M}_1 : \theta \neq 0$.
- Data D.

By Bayes' rule ($i = 0, 1$):

$$p(\mathcal{M}_i|D) = \frac{p(\mathcal{M}_i)p(D|\mathcal{M}_i)}{p(\mathcal{M}_0)p(D|\mathcal{M}_0) + p(\mathcal{M}_1)p(D|\mathcal{M}_1)}.$$

Then

$$\underbrace{\frac{p(\mathcal{M}_0|D)}{p(\mathcal{M}_1|D)}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}}_{\text{prior odds}} \times \underbrace{\frac{p(D|\mathcal{M}_0)}{p(D|\mathcal{M}_1)}}_{\text{Bayes factor, } BF_{01}} .$$

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

- Typical interpretation, e.g., $BF_{01} = 5$:

  The data are five times more likely to have occurred under $\mathcal{M}_0$ than under $\mathcal{M}_1$.

  or, alternatively,

  For any given prior odds, the posterior odds are five time more in favor of $\mathcal{M}_0$.

- $BF_{01} \in [0, \infty)$:
  - $BF_{01} < 1 \longrightarrow$ Support for $\mathcal{M}_1$ over $\mathcal{M}_0$.
  - $BF_{01} = 1 \longrightarrow$ Equal support for either model.
  - $BF_{01} > 1 \longrightarrow$ Support for $\mathcal{M}_0$ over $\mathcal{M}_1$.

Bayes factor have been praised in many instances.[1,2,3,4,5]

Here we take a critical look at Bayes factors.

---

[1] Dienes (2011).
[2] Dienes (2014).
[3] Masson (2011).
[4] Vampaemel (2010).
[5] Wagenmakers et al. (2018).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

List of issues

1. Bayes factors can be hard to compute. →
2. Bayes factors are sensitive to within-model priors. →
3. Use of ‘default’ Bayes factors. →
4. Bayes factors are not posterior model probabilities. →
5. Bayes factors do not imply a model is probably correct. →
6. Qualitative interpretation of Bayes factors. →
7. Bayes factors test model *classes*. →
8. Bayes factors $\longleftrightarrow$ parameter estimation. →
9. Bayes factors favor point $\mathcal{M}_0$. →
10. Bayes factors favor $\mathcal{M}_a$. →
11. Bayes factors often agree with p-values. →

I will focus on *some* of the issues, for time purposes.
The remaining are left as extra slides at the end (but we can discuss them too!!).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Bayes factors are sensitive to within-model priors

- Very well known.[1,2,3,4,5]
- Due to fact that the likelihood function is averaged over the prior to compute the marginal likelihood under a model:

$$P(D|\mathcal{M}_i) = \int_{\Theta_i} p(D|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)d\theta.$$

Example: Bias of a coin[6]

- $\mathcal{M}_0 : \theta = .5$ vs $\mathcal{M}_1 : \theta \neq .5$
- Data: 60 successes in 100 throws.
- Four within-model priors; all $Beta(a, b)$.

Prior	BF_{10}	Lee & Wagenmakers (2014)
Approx. to Haldane's prior ($a = .05$, $b = .05$)	0.09	'Strong' evidence for $\mathcal{M}_0$
Jeffreys' prior ($a = .5$, $b = .5$)	0.60	'Anecdotal' evidence for $\mathcal{M}_0$
Uniform prior ($a = 1$, $b = 1$)	0.91	'Anecdotal' evidence for $\mathcal{M}_0$
An informative prior ($a = 3$, $b = 2$)	1.55	'Anecdotal' evidence for $\mathcal{M}_1$

[1] Kass (1993).
[2] Gallistel (2009).
[3] Vampaemel (2010).
[4] Robert (2016).
[5] Withers (2002).
[6] Liu and Aitkin (2008).

- Arbitrarily vague priors are not allowed because the null model would be invariably supported. So, in the Bayes Factor context, vague priors will predetermine the test result![1]
- However, counterintuitively, improper priors *might* work.[2]
- The problem cannot be solved by increasing sample size.[3,4,5]

This behavior of Bayes factors is in sharp contrast with estimation of posterior distributions.[6,7]

---

[1] Morey and Rouder (2011).
[2] Berger and Pericchi (2001).
[3] Bayarri et al. (2012).
[4] Berger and Pericchi (2001).
[5] Kass and Raftery (1995).
[6] Gelman, Meng, and Stern (1996).
[7] Kass (1993).

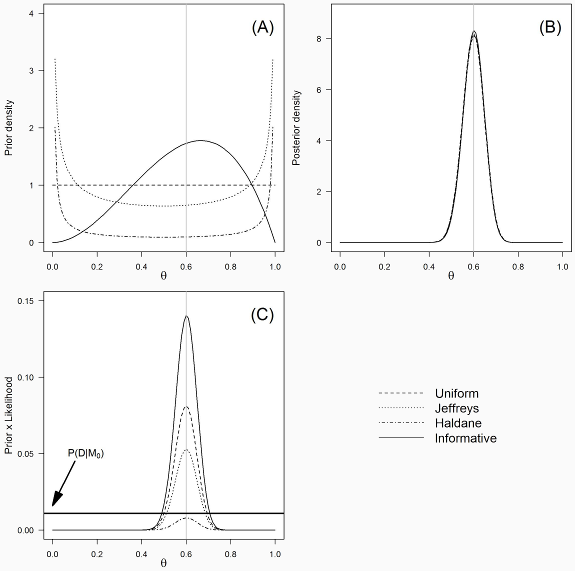


Figure 1: Data: 60 successes in 100 throws.

How to best choose priors then?

- Some defend informative priors should be part of model setup and evaluation.[1]
- Other suggest using default/ reference/ objective, well chosen, priors.[2,3,4,5]
- Perform sensitivity analysis.

[1] Vampaemel (2010).
[2] Bayarri et al. (2012).
[3] Jeffreys (1961).
[4] Marden (2000).
[5] Rouder et al. (2009).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Bayes factors are not posterior model probabilities

Bayes factors are not posterior model probabilities

Say that $BF_{01} = 32$; what does this mean?

After looking at the data, we revise our belief towards $\mathcal{M}_0$ by 32 times.

Q: What does this imply concerning the probability of each model, given the observed data?
A: On its own, nothing at all!

Bayes factors are the multiplicative factor converting prior odds to posterior odds. They say nothing directly about model probabilities.

$$\underbrace{\frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}}_{\text{prior odds}} \times \underbrace{\frac{p(D|\mathcal{M}_0)}{p(D|\mathcal{M}_1)}}_{\text{Bayes factor}} = \underbrace{\frac{p(\mathcal{M}_0|D)}{p(\mathcal{M}_1|D)}}_{\text{posterior odds}}$$

- Bayes factors say nothing about the plausability of each model in light of the data, that is, of $p(\mathcal{M}_i|D)$.
- Thus, Bayes factors $=$ rate of change of belief, not belief itself.[1]
- To compute $p(\mathcal{M}_i|D)$, prior model probabilities are needed:

$$p(\mathcal{M}_0|D) = \frac{\text{Prior odds} \times BF_{01}}{1 + \text{Prior odds} \times BF_{01}}, \quad p(\mathcal{M}_1|D) = 1 - p(\mathcal{M}_0|D).$$

Example

- Anna: Equal prior belief for either model.
- Ben: Strong prior belief for $\mathcal{M}_1$.
- $BF_{01} = 32$: Applies to Anna and Ben equally.

	$p(\mathcal{M}_0)$	$p(\mathcal{M}_1)$	BF_{01}	$p(\mathcal{M}_0\|D)$	$p(\mathcal{M}_1\|D)$	Conclusion
Anna	.50	.50	32	**.970**	.030	Favors $\mathcal{M}_0$
Ben	.01	.99		.244	**.756**	Favors $\mathcal{M}_1$

[1] Edwards, Lindman, and Savage (1963).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Bayes factors do not imply a model is probably correct

- A large Bayes factor, say, $BF_{10} = 100$, may mislead one to belief that $\mathcal{M}_1$ is true or at least more useful.
- Bayes factors are only a measure of relative plausibility among two competing models.
- $\mathcal{M}_1$ might actually be a dreadful model (e.g., lead to horribly wrong predictions), but simply less dreadful than its alternative $\mathcal{M}_0$.[1]
- Bayes factors provide no absolute evidence supporting either model under comparison.[2]
- Little is known as to how Bayes factors behave under model misspecification (but see[3]).

In general, I suggest:

- Avoid thinking about truth / falsehood.
- Instead, think about evidence in favor / against of a model.
- Bayes factors can indeed assist with this.

[1] Rouder (2014).

[2] Gelman and Rubin (1995).

[3] Ly, Verhagen, and Wagenmakers (2016).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Bayes factors $\longleftrightarrow$ parameter estimation

- Frequentist two-sided significance tests and confidence intervals (CIs) are directly related:
  The null hypothesis is rejected iff the null point is outside the CI.
- This is not valid in the Bayesian framework.[1]

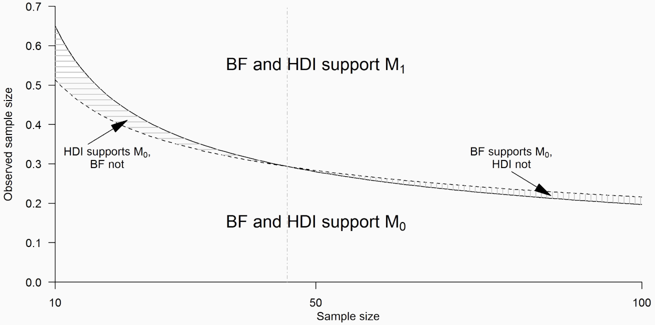


Figure 2: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{M}_0 : \mu = 0$ vs $\mathcal{M}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

[1] Kruschke and Liddell (2018b).

- There are many 'credible intervals', thus perhaps not surprising.
- Estimation and testing seem apart in the Bayesian world. Some argue they address different research questions[1,2,3,4] , but not everyone agrees.[5,6]

In particular, myself and Henk Kiers have recently argued that a unified Bayesian framework for testing and estimation is possible (Part 2 of today's talk).

[1] Kruschke (2011).
[2] Ly, Verhagen, and Wagenmakers (2016).
[3] Wagenmakers et al. (2018).
[4] Kruschke and Liddell (2018a).
[5] Robert (2016).
[6] Bernardo (2012).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Bayes factors favor point $\mathcal{M}_0$

- NHST is strongly biased against the point null model $\mathcal{M}_0$.[1,2,3,4]
- In other words, $p(\mathcal{M}_0|D)$ and p-values do not agree.
  (Yes, they are conceptually different![5])
- The discrepancy worsens as the sample size increases.

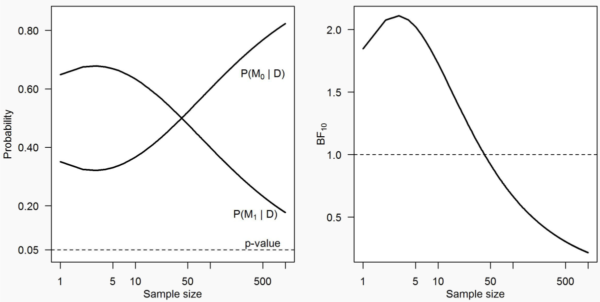


Figure 3: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{M}_0 : \mu = 0$ vs $\mathcal{M}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

[1] Edwards, Lindman, and Savage (1963).
[2] Dickey (1977).
[3] Berger and Sellke (1987).
[4] Sellke, Bayarri, and Berger (2001).
[5] Gigerenzer (2018).

- In this example, for $n > 42$ one rejects $\mathcal{M}_0$ under NHST whereas $BF_{10} < 1$ (indicating support for $\mathcal{M}_0$).
- In sum: Bigger ESs are needed for Bayes factor to sway towards $\mathcal{M}_1$. But, how much bigger?

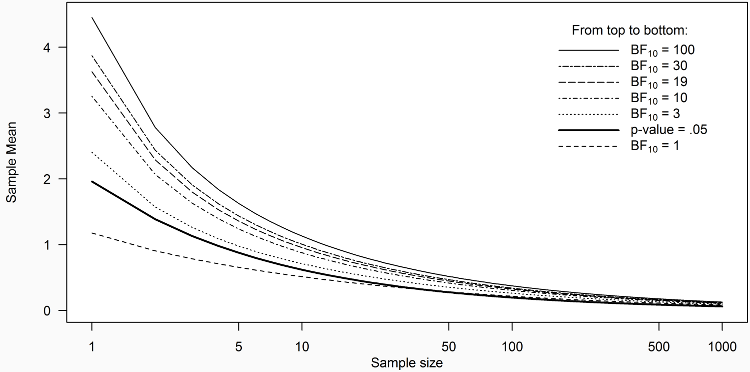


Figure 4: ESs required by BF_{10}, based of Jeffreys (1961) taxonomy.

Calibrate Bayes factors $\longleftrightarrow$ p-values?[1,2]

[1] Wetzels et al. (2011).

[2] Jeon and De Boeck (2017).

- Surprisingly, the previous result does not hold for one-sided $\mathcal{M}_0$ (e.g., comparing $\mu > 0$ and $\mu < 0$).[1,2]
- In this case, $p(\mathcal{M}_0|D)$ and p-values can be very close under a wide range of priors.

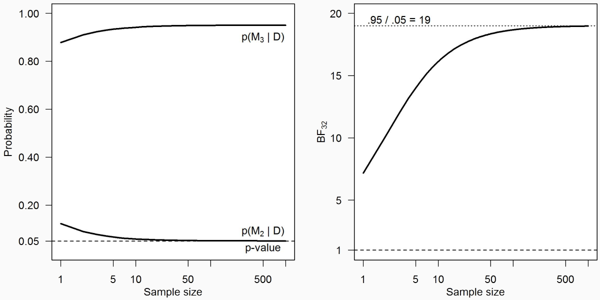


Figure 5: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{M}_2 : \mu \sim N^+(0, \sigma_1^2 = 1)$ vs $\mathcal{M}_3 : \mu \sim N^-(0, \sigma_1^2 = 1)$.

[1] Pratt (1965).

[2] Casella and Berger (1987).

Tuning just-significant ESs with Bayes factors:

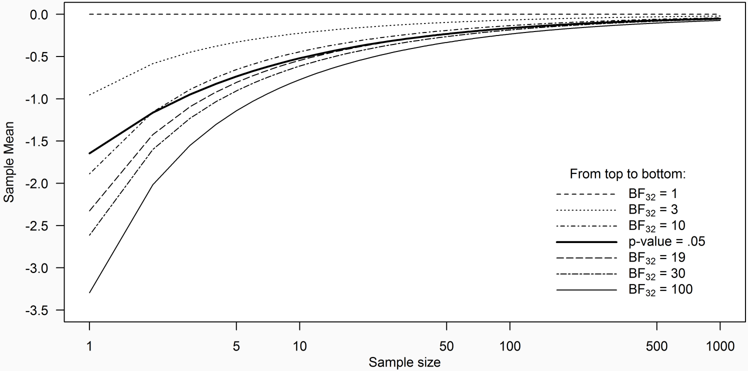


Figure 6: ESs required by BF_{10}, based of Jeffreys (1961) taxonomy.

- $p(\mathcal{M}_0|D)$ can be equal or even smaller than the p-value.[1]
- 'p-values overstate evidence against $\mathcal{M}_0$' $\longrightarrow$ Not always.[2]

Who to blame for this state of affairs?

We suggest the nature of the point null hypothesis; we are not alone.[3,4]
But others have argued in favor point of null hypotheses.[5,6,7,8,9,10]

'True' point hypotheses, really?![11,12,13]

[1] Casella and Berger (1987).
[2] Jeffreys (1961).
[3] Casella and Berger (1987).
[4] Vardeman (1987).
[5] Berger and Delampady (1987).
[6] Kass and Raftery (1995).
[7] Gallistel (2009).
[8] Konijn et al. (2015).
[9] Marden (2000).
[10] Morey and Rouder (2011).
[11] Berger and Delampady (1987).
[12] Cohen (1994).
[13] Morey and Rouder (2011).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Bayes factors favor $\mathcal{M}_a$

- Unless $\mathcal{M}_0$ is exactly true, $n \to \infty \implies BF_{01} \to 0$.
- Thus, both BF_{01} and the p-value approach 0 as n increases.
- It has be argued that this is a good property of Bayes factors (they are information consistent).[1]
- However, BF_{01} does ignore 'practical significance', or magnitude of ESs.[2]
- Meehl's paradox: For true negligible non-zero ESs, data accumulation should make it easier to reject a theory, not confirm it.[3,4]

---

[1] Ly, Verhagen, and Wagenmakers (2016).
[2] Morey and Rouder (2011).
[3] Meehl (1967).
[4] Kruschke and Liddell (2018b).

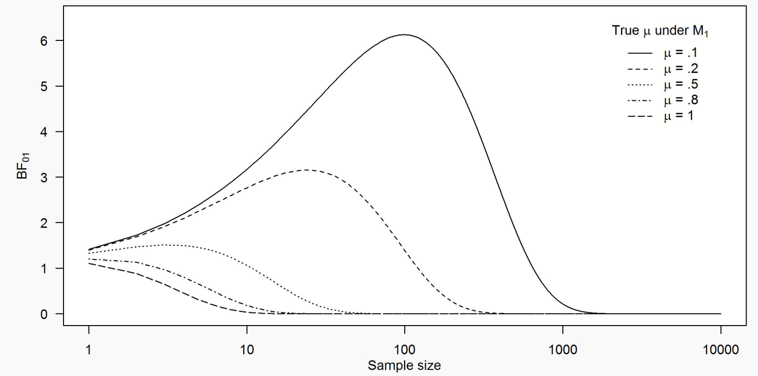


Figure 7: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{M}_0 : \mu = 0$ vs $\mathcal{M}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

- Consider $\mathcal{M}_0 : \theta = \theta_0$ vs $\mathcal{M}_0 : \theta \neq \theta_0$.
- As $n \to \infty$, Bayes factors accumulate evidence in favor of true $\mathcal{M}_1$ much faster than they accumulate evidence in favor of true $\mathcal{M}_0$.
- I.e., although Bayes factors allow drawing support for either model, they do so asymmetrically.[1]

[1] Johnson and Rossell (2010).

Bayes factors favor $\mathcal{M}_a$, II

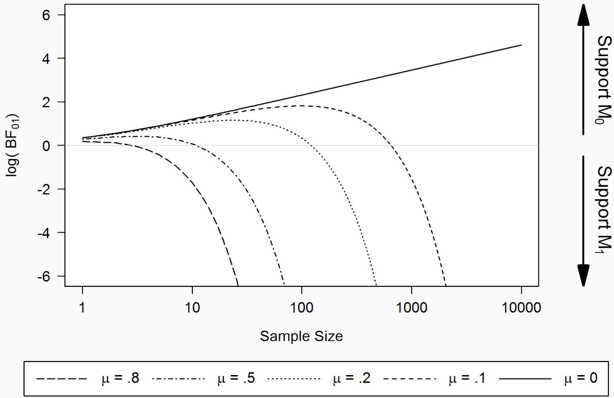


Figure 8: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{M}_0 : \mu = 0$ *vs* $\mathcal{M}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Bayes factors and the replication crisis

- It is increasingly difficult to ignore the current crisis of confidence in psychological research.
- Several key papers and reports made the ongoing state of affairs unbearable.[1,2,3,4,5,6]
- Some attempts to mitigate the problem have been put forward, including pre-registration and recalibration.[7,8]
- Some have suggested that a shift towards Bayesian testing is welcome.[9,10,11]

Would Bayes factors contribute to improving things?

---

[1] Ioannidis (2005).
[2] Simmons, Nelson, and Simonsohn (2011).
[3] Bem (2011).
[4] Wicherts, Bakker, and Molenaar (2011).
[5] John, Loewenstein, and Prelec (2012).
[6] Open Science Collaboration (2015).
[7] Benjamin et al. (2018).
[8] Lakens et al. (2018).
[9] Vampaemel (2010).
[10] Konijn et al. (2015).
[11] Dienes (2016).

What Bayes factors promise to offer might not be what researchers and journals are willing to use.[1]

- It has not yet been shown that the Bayes factors' ability to draw support for $\mathcal{M}_0$ will alleviate the bias against publishing null results ("lack of effects" are still too unpopular).
  Bayes factors need not be aligned with current publication guidelines.
- 'B-hacking'[2] is still entirely possible. New QRPs lurking around the corner?

[1] Savalei and Dunn (2015). [2] Konijn et al. (2015).

Part 1 – A Review of Issues About Null Hypothesis Bayesian Testing

Discussion

We think that:

- The use, abuse, and misuse of NHST and p-values is problematic. The statistical community is aware of this.[1]
- Bayes factors are an interesting alternative, but they do have limitations of their own.
- In particular, Bayes factors are also based on 'dichotomous modeling thinking': Given two models, which one is to be preferred?
  We favor a more holistic approach to model comparison.
- Bayes factors provide no direct information concerning effect sizes, their magnitude, and uncertainty.[2,3] This is sorely missed by this approach.

[1] Wasserstein and Lazar (2016).

[2] Wilkinson (1999).

[3] Kruschke and Liddell (2018a).

What to do?

- Truly consider whether testing is what you need.
- In particular, point hypotheses seem prone to trouble.
  How realistic are these hypotheses?
- Do estimation![1,2,3]
  Perform inference based on the entire posterior distribution. Report credible values. Compute posterior probabilities.

[1] Cohen (1994).

[2] Kruschke (2011).

[3] van der Linden and Chryst (2017).

Part 2 – With Bayesian estimation one can get all that Bayes factors offer, and more

Paper currently under revision.
Preprint here: `https://psyarxiv.com/zbpmy/`.

Part 2 – With Bayesian estimation one can get all that Bayes factors offer, and more

Motivation

- A link between NHBT and Bayesian estimation has been recently reiterated.[1]
- It requires the so-called spike-and-slab prior[2] :
    - A point mass probability on the null point.
    - A probability density function everywhere else.

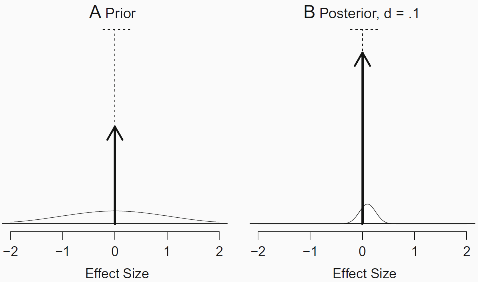


Figure 9: From Rouder et al. (2018). $\mathcal{M}_0 : \delta = 0$ vs $\mathcal{M}_1 : \delta \sim N(0, \sigma_0^2)$. $\delta = \frac{\mu}{\sigma} =$ std. ES.

[1] Rouder, Haaf, and Vandekerckhove (2018). [2] Mitchell and Beauchamp (1988).

Part 2 – With Bayesian estimation one can get all that Bayes factors offer, and more

Results

- We derived the closed-form expression of the posterior distribution based on the spike-and-slab prior.
- We show that the spike-and-slab prior can be approximated by a pure probability density function which we called the hill-and-chimney prior.
- We derived the closed-form expression of the posterior distribution based on the hill-and-chimney prior.
- We established that the hill-and-chimney prior converges to the spike-and-slab prior as the chimney's width converges to 0.
- The hill-and-chimney prior is not continuous. We offer an accurate approximation that is continuous, by means of mollification.[1]
- Importantly, Bayes factor values can be closely approximated by means of these posterior distributions based on (approx.) hill-and-chimney priors.
- Hence,

  With Bayesian estimation one can get all that Bayes factors offer, and more.

---

[1] Friedrichs (1944).

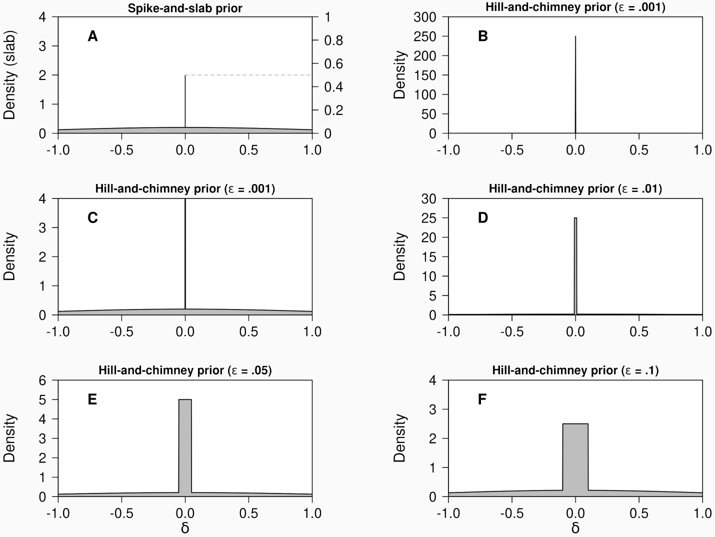


Figure 10: Spike-and-slab prior (**A**), hill-and-chimney prior (**B**–**F**).

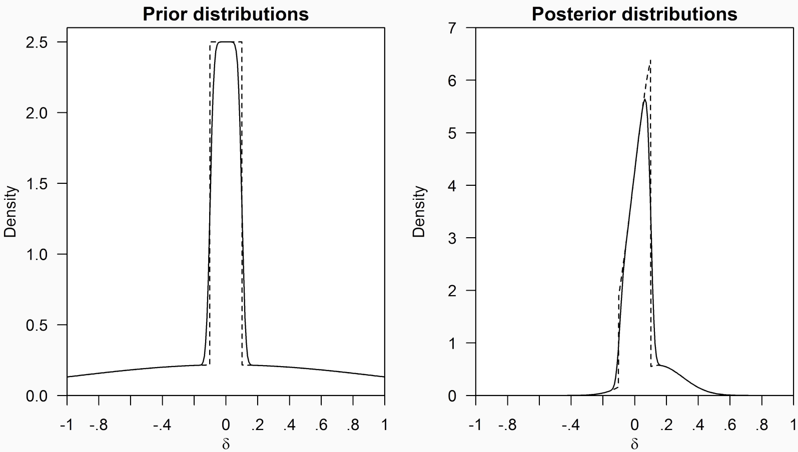
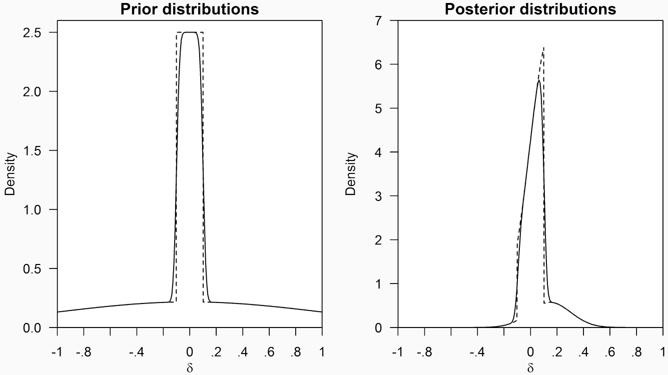


Figure 11: Approximating the hill-and-chimney prior by mollification ($n = 40$, $\delta = .15$, $\sigma = 1$, $\sigma_0 = 1$, $\varepsilon = .1$).

But 'what more' can Bayesian estimation offer?
$\longrightarrow$ Probabilities under the posterior distribution!



- $4.13 = BF_{01} \simeq$ posterior odds ratio $= \frac{P(\delta \in [-\varepsilon, \varepsilon] | \mathbf{y})}{P(\delta \notin [-\varepsilon, \varepsilon] | \mathbf{y})} = 3.81.$
- $P(\delta > 0 | \mathbf{y}) = .70.$
- $P(\delta > 0.1 | \mathbf{y}) = .18.$
- $P(\delta > 0.3 | \mathbf{y}) = .04.$

Part 2 – With Bayesian estimation one can get all that Bayes factors offer, and more

Discussion

- We fully integrated Bayesian testing and estimation for one simple model setting.
- The Bayes factor is only one of many possible probability statements under the posterior.
  So, estimation is much richer than testing.
- Spike-and-slab priors are difficult to justify.
  Hill-and-chimney priors are much more reasonable.
- Smooth continuous approximations to the hill-and-chimney prior work well.

Part 3 – A mathematical proof for optional stopping using NHBT

Paper almost ready to submit.

Part 3 – A mathematical proof for optional stopping using NHBT

Motivation

We focus on the optional stopping, or sequential testing, procedure to test between two models $\mathcal{M}_0 : \mu = \mu_0$ and $\mathcal{M}_1$ (e.g., $\mu = \mu_1$ or $\mu \neq \mu_0$):

1. Collect some data.
2. Perform the test.

   2a. Using NHST (choose α and n_{max} in advance):
   Compute p and...

   - ...if $p < \alpha$: STOP and retain $\mathcal{M}_1$.
   - ...if $p > \alpha$: Back to 1.

   Continue until either conclusive evidence or n_{max} is reached.

   2b. Using NHBT (choose BF_L, BF_U, and n_{max} in advance):
   Compute BF_{10} and...

   - ...if $BF_{10} < BF_L$: Stop and retain $\mathcal{M}_0$.
   - ...if $BF_{10} > BF_U$: Stop and retain $\mathcal{M}_1$.
   - ...if $BF_L < BF_{10} < BF_U$: Back to 1.

   Continue until either conclusive evidence or n_{max} is reached.

Optional stopping is a real problem under NHST.[1,2]

$\longrightarrow$ False positive rate $\gg \alpha$.

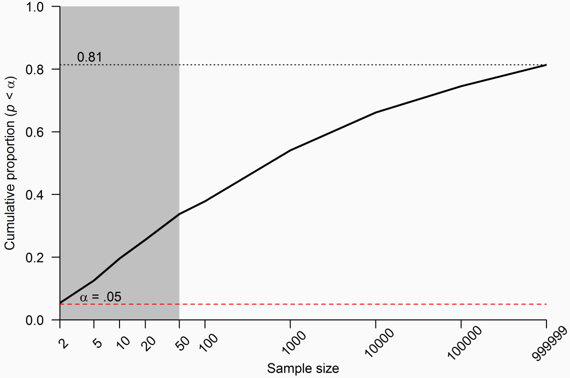


Figure 12: Proportion of false positives as a function of sample size under the frequentist optional stopping procedure, for a one-sample t-test.

[1] Armitage, McPherson, and Rowe (1969). [2] Jennison and Turnbull (1990).

What about using NHBT?

It has been argued through the years that optional stopping under the Bayesian paradigm is allowed.[1,2,3,4,5]

However, two recent papers disputed this state of affairs.[6,7]

Rouder offered a rebuttal to these papers in 2014.
(Title: ‘Optional stopping: No problem for Bayesians’).[8]

---

[1] Edwards, Lindman, and Savage (1963).
[2] Kass and Raftery (1995).
[3] Wagenmakers (2007).
[4] Wagenmakers et al. (2010).
[5] Francis (2012).
[6] Yu et al. (2014).
[7] Sanborn and Hills (2014).
[8] Rouder (2014).

Rouder claimed that Bayes factors are well calibrated under optional stopping.

The argument goes as follows:

- Assume prior odds equal to 1.
- This implies that
$$\underbrace{\frac{p(D|\mathcal{M}_1)}{p(D|\mathcal{M}_0)}}_{\text{Bayes factor, } BF_{10}} = \underbrace{\frac{p(\mathcal{M}_1|D)}{p(\mathcal{M}_0|D)}}_{\text{posterior odds}}.$$
- By definition of posterior odds, for any given value BF_{10},
  $\mathcal{M}_1$ is BF_{10} times more likely than $\mathcal{M}_0$ after considering the data.
- Rouder made two assertions:
  1. For any given value BF_{10},
     $\mathcal{M}_1$ is BF_{10} times more likely than $\mathcal{M}_0$ to have generated BF_{10}.
  2. The above statement also holds under optional stopping.

Part 3 – A mathematical proof for optional stopping using NHBT

Results

Rouder used simulations only to make his point, for two tests on the mean μ of a normal distribution with know variance σ^2:

- $\mathcal{M}_0 : \mu = 0$ *versus* $\mathcal{M}_1 : \mu = \mu_1$.
- $\mathcal{M}_0 : \mu = 0$ *versus* $\mathcal{M}_1 : \mu \sim \mathcal{N}(0, \sigma_1^2)$ with σ_1^2 known.

In our paper, we offer mathematical derivations to both of Rouder's assertions, for both tests above:

- We fully proved assertion 1 for both tests, for a fixed sample size n.
- We provide a proof of assertion 2 for a particular situation:
  After exactly one step of the optional stopping procedure.

Our trick:

We computed the sampling distribution of (the log of the) BF_{10} under $\mathcal{M}_0$ and $\mathcal{M}_1$, and showed that their ratio equals the BF_{10} itself.

Example: $\mathcal{M}_0 : \mu = 0$ *versus* $\mathcal{M}_1 : \mu = \mu_1$, with σ^2 known.

Bayes factor formula:

$$BF_{10} = \exp\left[\frac{n\mu_1(2\overline{X} - \mu_1)}{2\sigma^2}\right].$$

We worked with logarithms:

$$\ln(BF_{10}) = \frac{n\mu_1(2\overline{X} - \mu_1)}{2\sigma^2}.$$

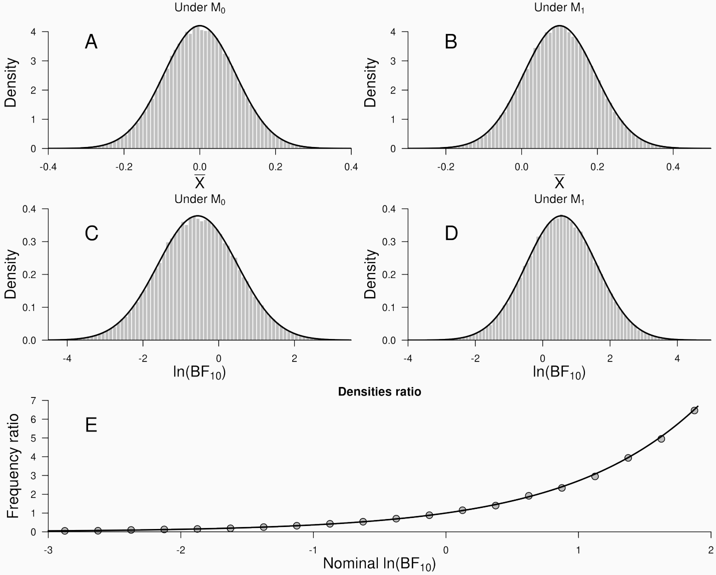


Figure 13: After $n = 10$ observations, with $\sigma = .3$ and $\mu_1 = .1$.

Part 3 – A mathematical proof for optional stopping using NHBT

Discussion

We offer a mathematical proof to a Bayes factor property suggested by Rouder (2014).

Is this conclusive evidence that Bayesian optional stopping is allowed?
Well, not just yet.[1]

However, in a very recent reply, Rouder again disagrees...
https://psyarxiv.com/m6dhw/

To be continued...

---

[1] Heide and Grünwald (2017).

Conclusion

I have spent some time learning about Bayes factors.

What do I now think of them?

I think that:

- Model comparison (including hypothesis testing) has a time and place in Psychology.
- However, and clearly, people test way too much.
- Model comparison says very little (nothing?) about how well a model fits to data.
- Testing need not be a prerequisite for estimation, unlike what some advocate.[1]
- Estimation quantifies uncertainty in ways that Bayes factors simply can not.
- Estimate ESs (direction, magnitude). Bayes factors ignore this!
- Avoid the dichotomous reasoning subjacent to Bayes factors.
- Bayes factors can be very useful (I use them!), but they should not *always* be the end of our inference.

[1] Wagenmakers et al. (2018).

Thank you!

Extra – Part 1

Extra – Part 1

Bayes factors can be hard to compute

$$BF_{01} = \frac{P(D|\mathcal{M}_0)}{P(D|\mathcal{M}_1)}.$$

Bayes factors are ratios of marginal likelihoods:

$$P(D|\mathcal{M}_i) = \int_{\Theta_i} p(D|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)d\theta$$

- The marginal likelihoods, $P(D|\mathcal{M}_i)$, are hard to compute in general.
- Resort to (not straightforward) numerical procedures[1,2]
- Alternatively, use software with prepackaged default priors and data models[3,4] (limited to specific models).

But: See bridge sampling by Quentin Gronau.

[1] Chen, Shao, and Ibrahim (2000).
[2] Gamerman and Lopes (2006).
[3] JASP Team (2018).
[4] Morey and Rouder (2018).

Extra – Part 1

Use of ‘default’ Bayes factors

- Priors matter a lot for Bayes factors.
- ‘Objective’ bayesians advocate using predefined priors for testing.[1,2,3]
- Albeit convenient, default priors lack empirical justification.[4]
- ‘Objective priors’ were derived under strong requirements[5,6] , which impose strong restrictions on the priors (“appearance of objectivity”[7]).
- Defaults are only useful to the extent that they adequately translate one’s beliefs.[8,9]
- Some default priors, like the now famous JZS prior[10,11,12] , still require a specification of a scale parameter. Its default value has also changed over time.[13,14]

---

1 Jeffreys (1961).
2 Berger and Pericchi (2001).
3 Rouder et al. (2009).
4 Robert (2016).
5 Bayarri et al. (2012).
6 Berger and Pericchi (2001).
7 Berger and Pericchi (ibid.).
8 Kruschke (2011).
9 Kruschke and Liddell (2018a).
10 Jeffreys (1961).
11 Zellner and Siow (1980).
12 Rouder et al. (2009).
13 Rouder et al. (ibid.).
14 Morey and Rouder (2018).

Extra – Part 1

Qualitative interpretation of Bayes factors

- Bayes factors are a continuous measure of evidence in $[0, \infty)$:
    - $BF_{01} > 1$: Data are more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$.
      The larger BF_{01}, the stronger the evidence for $\mathcal{M}_0$ over $\mathcal{M}_1$.
    - $BF_{01} < 1$: Data are more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$.
      The smaller BF_{01}, the stronger the evidence for $\mathcal{M}_1$ over $\mathcal{M}_0$.
- But, how 'much more' likely?
- Answer is not unique: Qualitative interpretations of strength are subjective (what is weak?, moderate?, strong?).[1,2,3,4]

This is not a problem of Bayes factor per se, but of practitioners requiring qualitative labels for test results.

[1] Jeffreys (1961).
[2] Kass and Raftery (1995).
[3] Lee and Wagenmakers (2013).
[4] Dienes (2016).

Extra – Part 1

Bayes factors test model *classes*

Consider testing $\mathcal{M}_0 : \theta = \theta_0$ vs $\mathcal{M}_1 : \theta \neq \theta_0$. Then

$$B_{01} = \frac{p(D|\mathcal{M}_0)}{p(D|\mathcal{M}_1)}, \quad \text{with} \quad p(D|\mathcal{M}_1) = \int p(D|\theta, \mathcal{M}_1)p(\theta|\mathcal{M}_1)d\theta.$$

- $p(D|\mathcal{M}_1)$ is a weighted likelihood for a model class:
  Each parameter value θ defines one particular model in the class.
- Bayes factors as ratios of likelihoods of model classes.[1]
- E.g., $BF_{01} = 1/5$: The data are five times more likely under the model class under $\mathcal{M}_1$, averaged over its prior distribution, than under $\mathcal{M}_0$.
- Catch: *The most likely model class need not include the true model that generated the data.*
  I.e., the Bayes factor may fail to indicate the class that includes the data-generating model (in case it exists, of course).[2]

---

[1] Liu and Aitkin (2008).

[2] Liu and Aitkin (ibid.).

Extra – Part 1

Bayes factors often agree with p-values

p-values are often accused of being 'violently biased against the null hypothesis'.[1,2]

But this is not always true.[3]

Trafimow's argument:
Consider $p(D|\mathcal{M}_1)$, i.e., the likelihood of the observed data under the *alternative* model.

$$p(\mathcal{M}_0|D) = \frac{p(\mathcal{M}_0)p(D|\mathcal{M}_0)}{p(\mathcal{M}_0)p(D|\mathcal{M}_0) + [1 - p(\mathcal{M}_0)]p(D|\mathcal{M}_1)}$$

Suppose p is small (say, $< .05$).

- If $p(D|\mathcal{M}_1)$ is very small then $p(\mathcal{M}_0|D)$ is close to 1 for $p(D|\mathcal{M}_0)$ fixed.
  Disagreement with p.
- But, if $p(D|\mathcal{M}_1)$ is large then $p(\mathcal{M}_0|D)$ is small.
  Agreement with p.

[1] Edwards (1965).

[2] Wagenmakers et al. (2018).

[3] Trafimow (2003).

Conclusion:
When data are more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$, Bayes factors and p-values tend to agree with each other.

The p-value, by definition, is oblivious to the likelihood of the data under $\mathcal{M}_1$.

This is why the p-value is sometimes biased against $\mathcal{M}_0$.

NHBT allows drawing support for $\mathcal{M}_0$, unlike NHST.

So, large p-values cannot be used as evidence in favor of $\mathcal{M}_0$, but large BF_{01} values can.