The Bayes factor

Statistical properties and a diagnosis of its use in applied research

Jorge N. Tendeiro
tendeiro@hiroshima-u.ac.jp
www.jorgetendeiro.com

Slides at www.jorgetendeiro.com/talks/2023_HU_slides.pdf

06 January, 2023

Hiroshima University

1. NHST and its shortcomings (quickly).
2. Introduction to the Bayes factor.
3. Properties of the Bayes factor.
4. X-ray of applications of Bayes factors in the literature.

Jorge N. Tendeiro
Hiroshima University

Officially:
Office of Research and Academia-Government-Community Collaboration

For research:
Education and Research Center for Artificial Intelligence and Data Innovation

Research interests:

- Three-mode component analysis (PhD).
- Item response theory, namely person-fit analysis.
- Bayesian inference, namely the Bayes factor.
- Various types of statistical modeling through collaborations.

1. NHST and its shortcomings (quickly)

By *null hypothesis significance testing* (NHST), I am referring to the blend[1,2] between

Fisher's *significance testing*

and

Neyman and Pearson's *hypothesis testing*.

[1] Lehmann (1993).

[2] Hubbard (2004).

Misconceptions concerning NHST and its infamous p-value (and also the confidence interval) are well documented in the literature.[1,2,3,4,5,6,7]

Various science fields are experiencing a crisis of confidence, as many researchers believe published results are not as well supported as claimed.

Q: Why?
A: Among several other reasons (QRPs[8,9]), due to overreliance on, and misuse of NHST and p-values.[10,11,12,13]

[1] Belia et al. (2005).
[2] Falk and Greenbaum (1995).
[3] Goodman (2008).
[4] Greenland et al. (2016).
[5] Haller and Kraus (2002).
[6] Hoekstra et al. (2014).
[7] Oakes (1986).
[8] John, Loewenstein, and Prelec (2012).
[9] Simmons, Nelson, and Simonsohn (2011).
[10] Edwards, Lindman, and Savage (1963).
[11] Cohen (1994).
[12] Nickerson (2000).
[13] Wagenmakers (2007).

Here is a short, not exaustive, list:[1,2]

- $p =$ probability of $\mathcal{H}_0$ being true.
- $p < \alpha \implies \mathcal{H}_0$ is false.
- $p > \alpha \implies \mathcal{H}_0$ is true.
- $p > \alpha \implies \mathcal{H}_0$ is likely true.
- Relation between p and effect sizes.
- $p =$ probability of observed data under $\mathcal{H}_0$.
- $p < \alpha \implies$ the probability of a type I error is α.
- Statistically significant $\simeq$ practically significant.
- $p > \alpha \implies$ effect size is small.
- ...

[1] Goodman (2008).

[2] Greenland et al. (2016).

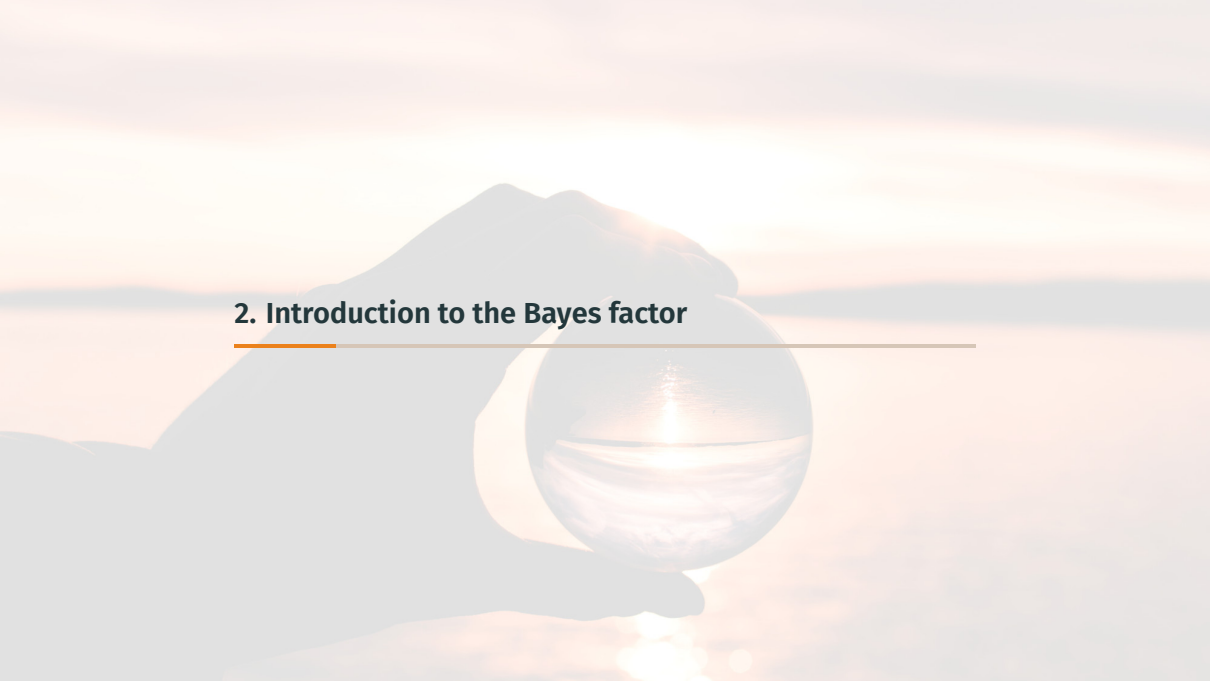
Is the p-value an uninteresting probability?

$$p = P\Big[\underbrace{\text{observed data (or more extreme)}}_{\text{data}} \,|\, \underbrace{\mathcal{H}_0}_{\text{theory}}\Big].$$

Arguably, researchers care more about the reversed conditional probability:

$$P(\text{theory}|\text{data}).$$

This leads us to the Bayes factor (well, only *kind of*).



2. Introduction to the Bayes factor

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

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

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

- Two hypotheses (or models) to compare, for instance $\mathcal{H}_0 : \theta = 0$ *vs* $\mathcal{H}_1 : \theta \neq 0$.
- Data D.

Assume that either $\mathcal{H}_0$ or $\mathcal{H}_1$ must hold true.
Then by Bayes' rule ($i = 0, 1$):

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

and dividing member by member leads to

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

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

$$\boxed{BF_{01} = \frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)}}$$

For instance, $BF_{01} = 5$:

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

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

For instance, $BF_{01} = 5$:

After observing the data, my relative belief in $\mathcal{H}_0$ over $\mathcal{H}_1$ increased 5 times.

This holds regardless of the initial relative belief of a rational agent:

Prior belief in...		Prior odds	$\mathbf{BF}_{01}$	Posterior odds	Posterior belief on...	
$\mathcal{H}_0$	$\mathcal{H}_1$				$\mathcal{H}_0$	$\mathcal{H}_1$
$1/2 = .50$	$1/2 = .50$	1	5	5	$5/6 = .83$	$1/6 = .17$
$2/3 = .67$	$1/3 = .33$	2	5	10	$10/11 = .91$	$1/11 = .09$
$1/10 = .01$	$9/10 = .90$	1/9	5	5/9	$5/14 = .36$	$9/14 = .64$

$BF_{01} = \frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)} \in [0, \infty)$:

- $BF_{01} > 1 \longrightarrow$ Evidence in favor of $\mathcal{H}_0$ over $\mathcal{H}_1$.
- $BF_{01} = 1 \longrightarrow$ Equal support for either model.
- $BF_{01} < 1 \longrightarrow$ Evidence in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$.

Some qualitative cutoff labels have been suggested, for instance[1,2,3] .

Here's Kass and Raftery's classifier:

$\mathbf{BF}_{01}$	**Strength of evidence in favor of $\mathcal{H}_0$**
$1 - 3$	Not worth more than a bare mention
$3 - 20$	Positive
$20 - 150$	Strong
> 150	Very strong

For $BF_{01} < 1$, use $BF_{10} = \frac{1}{BF_{01}}$ as strength of evidence in favor of $\mathcal{H}_1$.

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

[3] Lee and Wagenmakers (2013).

$$BF_{01} = \frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)}$$

For simpler models there are a few R packages available to assist with the computations:

- `BayesFactor`[1] (mostly used).
- `bain`.[2]
- `easystats`.[3]
- `bayestestR`.[4]
- `brms`[5] and `rstanarm`,[6] relying on the `bridgesampling`[7] package.

There is also JASP, a handy and open source GUI.

[1] Morey and Rouder (2022).
[2] Gu et al. (2021).
[3] Lüdecke et al. (2022).
[4] Makowski, Ben-Shachar, and Lüdecke (2019).
[5] Bürkner (2021).
[6] Goodrich et al. (2022).
[7] Gronau, Singmann, and Wagenmakers (2020).

$$\boxed{BF_{01} = \frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)}}$$

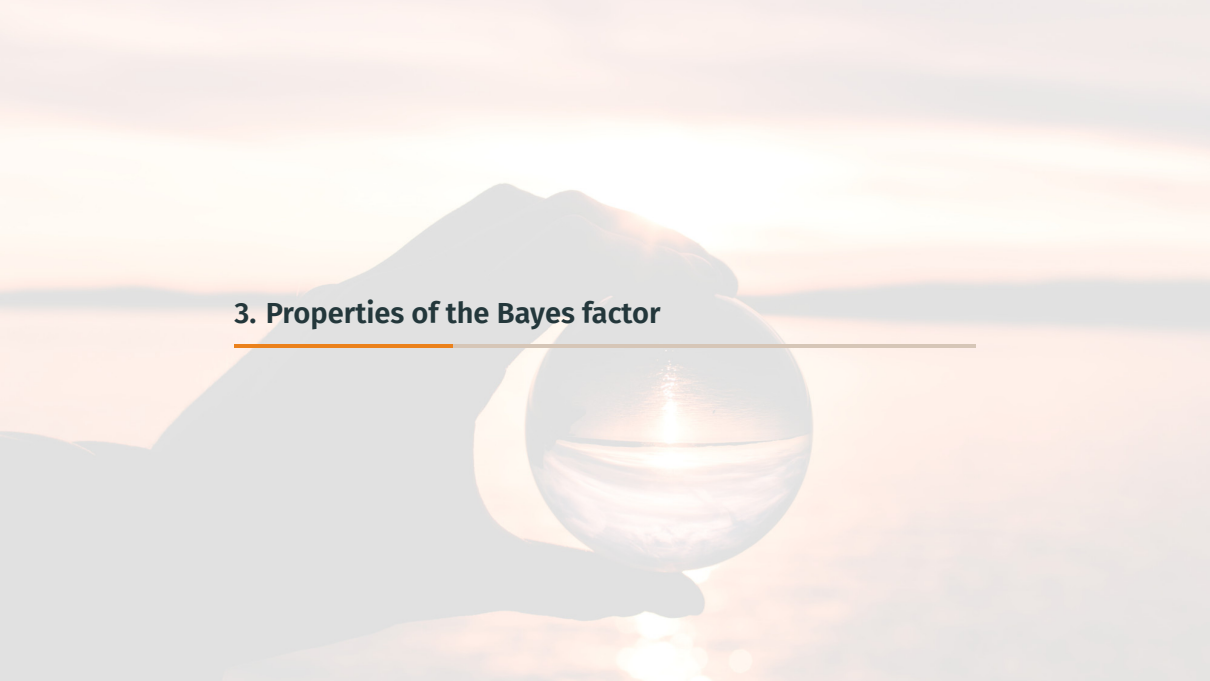
Essentially, any two statistical models that make predictions are in theory eligible to be compared via the Bayes factor.

We "just" need to evaluate each model's marginal likelihood:

$$P(D|\mathcal{H}_i) = \int_{\Theta_i} \underbrace{p(D|\theta, \mathcal{H}_i)}_{\text{likelihood}} \underbrace{p(\theta|\mathcal{H}_i)}_{\text{prior}} d\theta.$$

There are various numerical procedures for this (e.g.,[1,2,3,4,5,6,7,8]), but recently bridge sampling has been of great practical use (in combination JAGS, Stan, or NIMBLE).

[1] Berger and Pericchi (2001).
[2] Carlin and Chib (1995).
[3] Chen, Shao, and Ibrahim (2000).
[4] Gamerman and Lopes (2006).
[5] Gelman and Meng (1998).
[6] Green (1995).
[7] Gronau et al. (2017).
[8] Kass and Raftery (1995).



3. Properties of the Bayes factor

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

But, surprisingly, I could not find many sources with critical appraisals of the Bayes factor.

I did exactly this a couple of years ago.[6]

[1] Dienes (2011).
[2] Dienes (2014).
[3] Masson (2011).
[4] Vanpaemel (2010).
[5] Wagenmakers et al. (2018).
[6] Tendeiro and Kiers (2019).

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 $\mathcal{H}_0$. → →
10. Bayes factors favor $\mathcal{H}_a$. → →
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!!).

3. Properties of the Bayes factor

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:

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

Example: Bias of a coin[6]

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

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

[1] Kass (1993).
[2] Gallistel (2009).
[3] Vanpaemel (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).

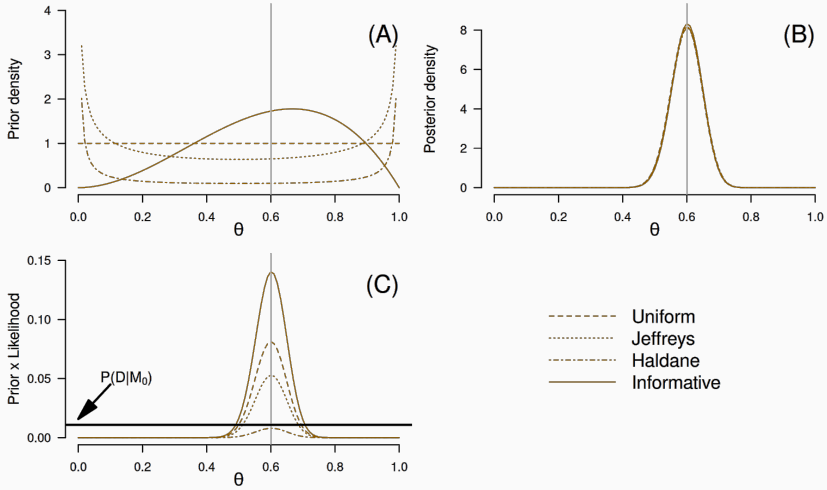


Figure 1: Data: 60 successes in 100 throws.

Bayes factors are sensitive to priors

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

3. Properties of the Bayes factor

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 $\mathcal{H}_0$ 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.

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

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

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

Example

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

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

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

3. Properties of the Bayes factor

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]

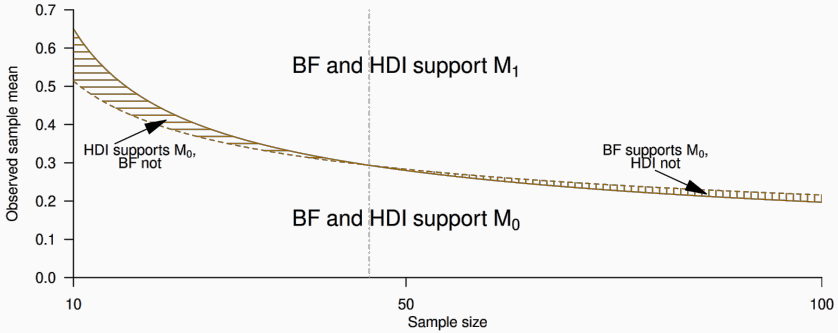


Figure 2: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{H}_0 : \mu = 0$ *vs* $\mathcal{H}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

[1] Kruschke and Liddell (2018a).

- 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 (https://psyarxiv.com/zbpmy/).[7]

[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).
[7] Tendeiro and Kiers (2022).

3. Properties of the Bayes factor

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

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

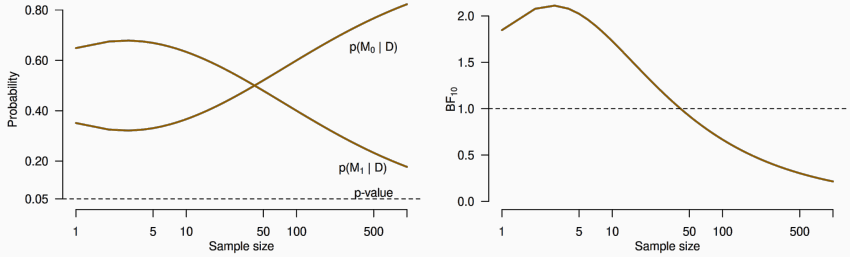


Figure 3: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{H}_0 : \mu = 0$ *vs* $\mathcal{H}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

[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 $\mathcal{H}_0$ under NHST whereas $BF_{10} < 1$ (indicating support for $\mathcal{H}_0$).
- In sum: Bigger ESs are needed for the Bayes factor to sway towards $\mathcal{H}_1$.
  But, how much bigger?

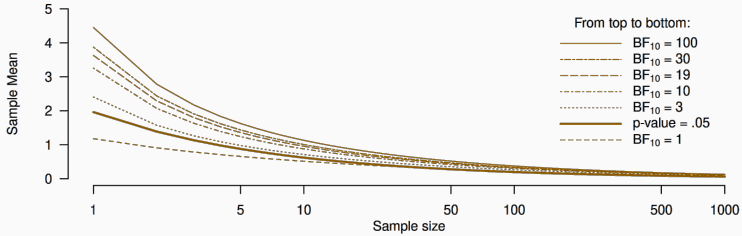


Figure 4: ESs required by BF_{10}, based of Jeffreys (1961) taxonomy.

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

[1] Wetzels et al. (2011).

[2] Jeon and De Boeck (2017).

3. Properties of the Bayes factor

Bayes factors favor $\mathcal{H}_a$

- Unless $\mathcal{H}_0$ 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 (2018a).

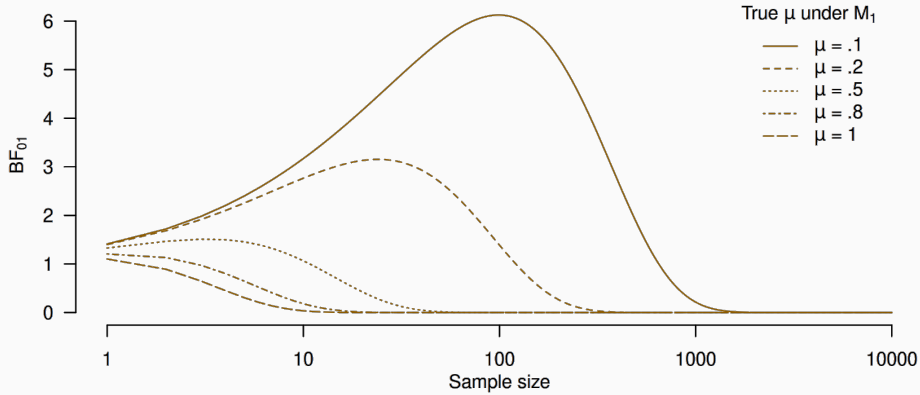


Figure 5: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{H}_0 : \mu = 0$ *vs* $\mathcal{H}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

3. Properties of the Bayes factor

Bayes factors and the replication crisis

- It is increasingly difficult to ignore the current crisis of confidence in 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] OSC (2015).
[7] Benjamin et al. (2018).
[8] Lakens et al. (2018).
[9] Vanpaemel (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 $\mathcal{H}_0$ 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).

3. Properties of the Bayes factor

Discussion

I 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?
  I 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 and Task Force on Statistical Inference (1999). [3] Kruschke and Liddell (2018b).

What to do?

- Truly consider whether testing is what is needed.
- 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).

4. X-ray of applications of Bayes factors in the literature

Until recently, there was no characterization of the use of the Bayes factor in applied research.

Wong and colleagues[1] were the first to start unveiling the current state of affairs.

In an ongoing effort, I am currently extending the work of Wong et al..
Here I report the details and main findings of my study.

Work with Henk Kiers, Rink Hoekstra, Tsz Keung Wong, and Richard Morey.

Preprint (under review):
https://psyarxiv.com/du3fc/

[1] Wong, Kiers, and Tendeiro (2022).

Background:
Social Sciences.

Target:
NHBT and the Bayes factor in particular.

Motivation:
Bayes factors have been regularly used since, say, 2010.
It is very recent.
Not many researchers have received formal training.
It is unclear how things are working out.

Google Scholar (2010−):

```
("bayes factor" AND "bayesian test" AND psychol)
```

Web of Science:

```
(TI=((bayes factor OR bayes* selection OR bayes* test*) AND psycho*) OR
AB=((bayes factor OR bayes* selection OR bayes* test* OR bf*) AND psychol*) OR
AK=((bayes factor OR bayes* selection OR bayes* test* OR bf*) AND psychol*))
AND PY=(2010-2022)
```

$109 + 58 = 167$ papers (after selection).

	Criterion	Brief description
QRIP	1 – Describing the BF as posterior odds	Defining or elaborating on BFs as posterior odds ratios.
	3a – Missing explanation for the chosen priors	The reason or justification for the chosen priors is not provided.
	3b – No mention to the priors used	It is unclear which priors were used under either model.
	3c – Incomplete info regarding the priors used	E.g., only providing the distribution family (“Cauchy”).
	4 – Not referring to the comparison of models	Presenting BFs as absolute evidence for one of the two models.
	5 – Making absolute statements	Based on the BF, concluding that there is (not) an effect.
	6 – Using BF as posterior odds	Interpreting BFs as ratios of posterior model probabilities.
	7 – Considering BF as effect size	Associating the size of the BF to the size of the effect.
	9 – Inconclusive evidence as evidence of absence	Stating that there is no effect when faced with inconclusive evidence.
	10 – Interpreting ranges of BF values only	Interpreting the Bayes factor simply using cutoffs (e.g., 1-3, 3-10).
Usage	A – Default prior	Justifying using a prior because it is ‘the’ default.
	B – Null results	Bayes factors as a follow-up to non-significant outcomes from NHST.
	C – Presence *versus* absence	Bayes factors to distinguish between the presence and the absence of an effect.

	Criterion	Count	Percentage
QRIP	1 – Describing the BF as posterior odds	22	13.2%
	3a – Missing explanation for the chosen priors	18	10.8%
	3b – No mention to the priors used	50	29.9%
	3c – Incomplete info regarding the priors used	10	6.0%
	4 – Not referring to the comparison of models	104	62.3%
	5 – Making absolute statements	59	35.3%
	6 – Using BF as posterior odds	34	20.4%
	7 – Considering BF as effect size	7	4.2%
	9 – Inconclusive evidence as evidence of absence	6	3.6%
	10 – Interpreting ranges of BF values only	9	5.4%
Usage	A – Default prior	59	35.3%
	B – Null results	27	16.2%
	C – Presence *versus* absence	30	18.0%

Overall:

- 149 papers (89.2%) displayed at least one QRIP.
- 104 papers (62.3%) displayed at least two QRIPs.

We reasoned over the reasons behind the found problems.

Below is a selected synopsis of our considerations.

4. X-ray of applications of Bayes factors in the literature

Bayes factors $\longleftrightarrow$ posterior odds

> *"The alternative hypothesis is 2 times more likely than the null hypothesis ($B_{+0} = 2.46$; Bayesian 95 % CI [0.106, 0.896])."*

Possible explanations:

- Principle of indifference.
- Overselling Bayes as the *theory of inverse probability*.[1]
- Cognitive dissonance.

[1] Jeffreys (1961).

4. X-ray of applications of Bayes factors in the literature

Dealing with priors

Reporting nothing at all (30%) or relying on software defaults (35%) was quite common.

Possible explanations:

- Lack of awareness.
- Economic writing style.
- Default priors to...
  ...ease comparison, avoid specification, 'objectivity'.

4. X-ray of applications of Bayes factors in the literature

Bayes factors as *relative* evidence

> *"With this 'stronger' VB05 prior, we found strong evidence for the null hypothesis (BFs_{null} ranging from 12.7 to 22.7 for the 5 ROIs)."*

Possible explanations:

- Writing style.
- Implicitly assumed.
- Increased impact.

4. X-ray of applications of Bayes factors in the literature

Bayes factors to establish absence/presence

> *"For 6-year-olds, there was no difference between environments (*$M_{smooth} = 2.11$ *vs.* $M_{rough} = 1.93$, $t(52) = 1.0$, $p = 0.31$, $d = 0.3$, $BF = .42$*)."*

Possible explanations:

- Increased impact.
- Avoid uncertainty.
- Writing style.
- Influence from NHST.
- Decision making.

What’s next?

A follow-up study is in preparation.

- Create and deploy a Shiny app that illustrates correct and incorrect usage of the Bayes factor.
- Assess the efficacy of this app by means of an experiment.

Conclusion

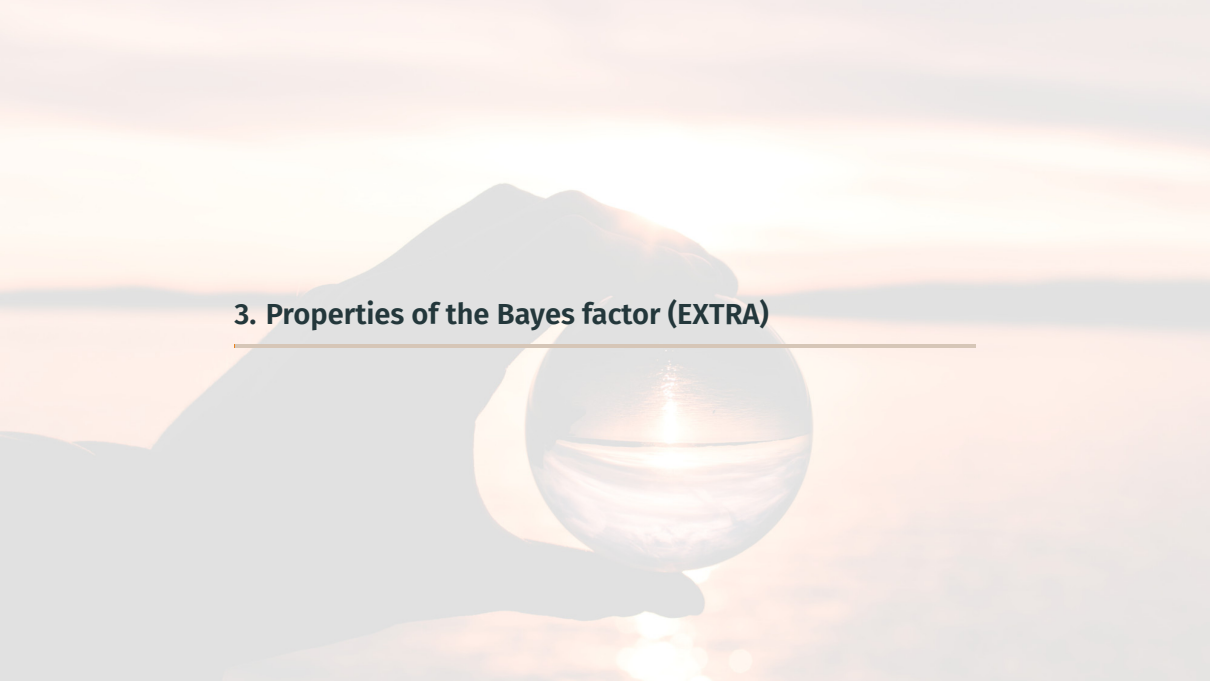
I have spent some time learning about Bayes factors.
What do I now think of them?

I think that:

- Model comparison (including hypothesis testing) is really important.
- However, and clearly, researchers 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).

Questions?



3. Properties of the Bayes factor (EXTRA)

3. Properties of the Bayes factor (EXTRA)

Bayes factors can be hard to compute

Bayes factors are hard to compute

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

Bayes factors are ratios of marginal likelihoods:

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

- The marginal likelihoods, $P(D|\mathcal{H}_i)$, 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 (2022).
[4] Morey and Rouder (2022).

3. Properties of the Bayes factor (EXTRA)

Use of ‘default’ Bayes factors

‘Default’ Bayes factors lack justification

- 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 (2018b).
[10] Jeffreys (1961).
[11] Zellner and Siow (1980).
[12] Rouder et al. (2009).
[13] Rouder et al. (ibid.).
[14] Morey and Rouder (2022).

3. Properties of the Bayes factor (EXTRA)

Bayes factors do not imply a model is probably correct

Bayes factors do not imply a model is correct

- A large Bayes factor, say, $BF_{10} = 100$, may mislead one to belief that $\mathcal{H}_1$ is true or at least more useful.
- Bayes factors are only a measure of relative plausibility among two competing models.
- $\mathcal{H}_1$ might actually be a dreadful model (e.g., lead to horribly wrong predictions), but simply less dreadful than its alternative $\mathcal{H}_0$.[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, it seems best to:

- 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).

3. Properties of the Bayes factor (EXTRA)

Qualitative interpretation of Bayes factors

Interpretation of Bayes factors can be ambiguous

- Bayes factors are a continuous measure of evidence in $[0, \infty)$:
  - $BF_{01} > 1$: Data are more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$.
    The larger BF_{01}, the stronger the evidence for $\mathcal{H}_0$ over $\mathcal{H}_1$.
  - $BF_{01} < 1$: Data are more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$.
    The smaller BF_{01}, the stronger the evidence for $\mathcal{H}_1$ over $\mathcal{H}_0$.
- 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).

3. Properties of the Bayes factor (EXTRA)

Bayes factors test model *classes*

Bayes factors test model *classes*

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

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

- $p(D|\mathcal{H}_1)$ 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 $\mathcal{H}_1$, averaged over its prior distribution, than under $\mathcal{H}_0$.
- 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.).

3. Properties of the Bayes factor (EXTRA)

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

Bayes factors don't favor one-sided $\mathcal{H}_0$

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

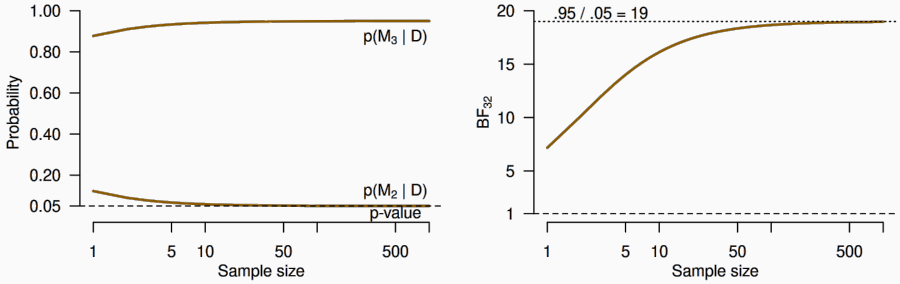


Figure 6: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{H}_2 : \mu \sim N^+(0, \sigma_1^2 = 1)$ *vs* $\mathcal{H}_3 : \mu \sim N^-(0, \sigma_1^2 = 1)$.

[1] Pratt (1965).

[2] Casella and Berger (1987).

Bayes factors don't favor one-sided $\mathcal{H}_0$

Tuning just-significant ESs with Bayes factors:

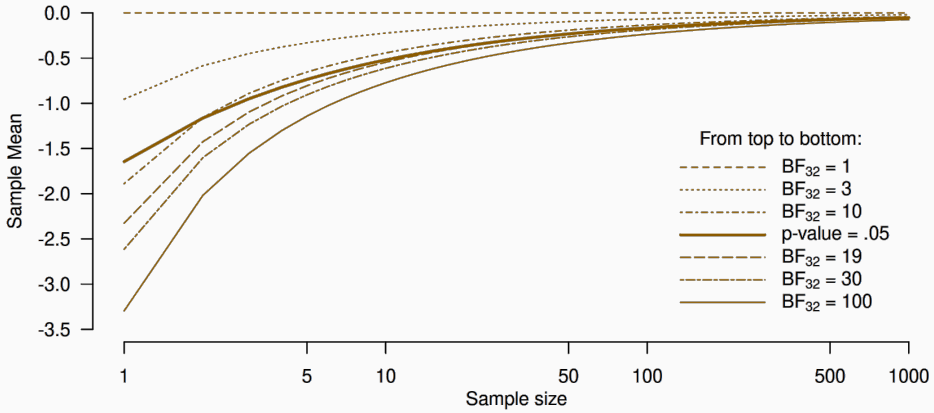


Figure 7: ESs required by BF_{32}, based of Jeffreys (1961) taxonomy.

Bayes factors don't favor one-sided $\mathcal{H}_0$

- $p(\mathcal{H}_0|D)$ can be equal or even smaller than the p-value.[1]
- 'p-values overstate evidence against $\mathcal{H}_0$' $\longrightarrow$ 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).

3. Properties of the Bayes factor (EXTRA)

Bayes factors favor $\mathcal{H}_a$

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

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

[1] Johnson and Rossell (2010).

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

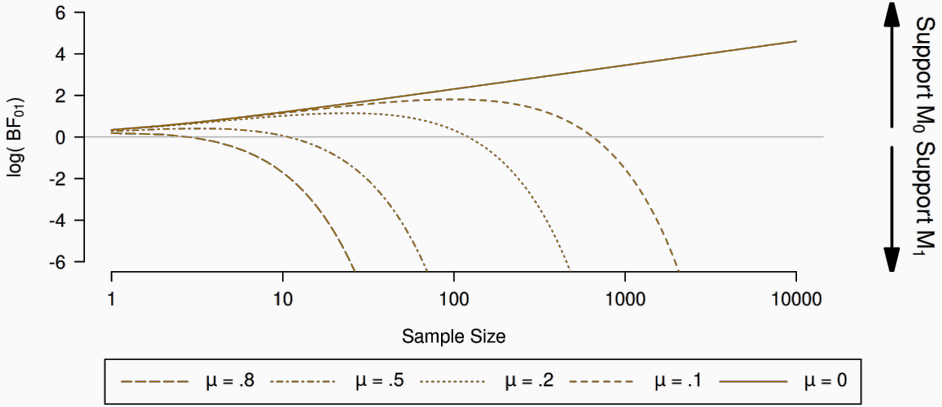


Figure 8: Data: $Y_i \sim N(\mu, \sigma^2 = 1)$. $\mathcal{H}_0 : \mu = 0$ *vs* $\mathcal{H}_1 : \mu \sim N(0, \sigma_1^2 = 1)$.

3. Properties of the Bayes factor (EXTRA)

Bayes factors often agree with p-values

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 $p(D|\mathcal{H}_1)$, i.e., the likelihood of the observed data under the *alternative* model.

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

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

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

[1] Edwards (1965). [2] Wagenmakers et al. (2018). [3] Trafimow (2003).

Bayes factors often agree with p-values

Conclusion:
When data are more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$, 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 $\mathcal{H}_1$.

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

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

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