Misuse of Bayesian inference in applied research

Current status and constructive feedback

Jorge N. Tendeiro
Hiroshima University
tendeiro@hiroshima-u.ac.jp

Slides and references:
https://www.jorgetendeiro.com/talk/2023_csp/

03 February, 2023

Today's talk

1. NHST and its shortcomings (briefly).
2. Introduction to the Bayes factor.
3. Properties of the Bayes factor.
4. The Bayes factors in applied research.
5. Conclusions, next steps.

1. NHST and its shortcomings (briefly)

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]

[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).

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?

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

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:

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

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.

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

[3] Etz and Vandekerckhove (2018).

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, $\mathcal{H}_0$ *vs* $\mathcal{H}_1$.
- Data D.

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

[3] Etz and Vandekerckhove (2018).

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, $\mathcal{H}_0$ *vs* $\mathcal{H}_1$.
- Data D.

Assuming that either $\mathcal{H}_0$ or $\mathcal{H}_1$ must hold true, then it can be shown that[3]

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

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

[3] Etz and Vandekerckhove (2018).

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

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 by 5 times.

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

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

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

This holds regardless of the initial relative belief (i.e., prior odds) of a rational agent.

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

- $BF_{01} > 1 \longrightarrow$ 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 $\mathcal{H}_0$.

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

[3] Lee and Wagenmakers (2013).

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

- $BF_{01} > 1 \longrightarrow$ 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).

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

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, that is, $p(D|\mathcal{H}_i)$ for $i = 0, 1$.

There are various numerical procedures for this[1,2,3,4,5,6,7,8] .

[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).

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

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).

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.

[1] Dienes (2011).
[2] Dienes (2014).
[3] Masson (2011).
[4] Vanpaemel (2010).
[5] Wagenmakers et al. (2018).
[6] Tendeiro and Kiers (2019).
[7] Tendeiro, Kiers, and Ravenzwaaij (2022).
[8] Tendeiro and Kiers (2023a).
[9] Tendeiro and Kiers (2023b).

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 have been doing this for a few years now.[6,7,8,9]

[1] Dienes (2011).
[2] Dienes (2014).
[3] Masson (2011).
[4] Vanpaemel (2010).
[5] Wagenmakers et al. (2018).
[6] Tendeiro and Kiers (2019).
[7] Tendeiro, Kiers, and Ravenzwaaij (2022).
[8] Tendeiro and Kiers (2023a).
[9] Tendeiro and Kiers (2023b).

1. Bayes factors are not posterior odds! →
2. Bayes factors are (at least *can be*) sensitive to priors! →
3. Bayes factors are a measure of relative evidence! →
4. Bayes factors can not establish absence/presence! →
5. Bayes factors are not an effect size measure! →
6. Inconclusive evidence is not evidence of absence! →
7. Bayes factors are a continuous measure of relative evidence! →

For the rest of this presentation, I will:

- Present the results of a study aiming at studying the occurrence of misconceptions in the literature.
- Explain each misconception.
- Speculate on why these misconceptions come about.

4. The Bayes factors in applied research

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.

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

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. The Bayes factors in applied research

Bayes factors are *not* posterior odds

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

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.

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

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!

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

- Bayes factors = rate of *change of* belief, not the *updated* belief.[1]
- The updated belief is captured by the posterior odds and posterior model probabilities.
- To compute these, prior model probabilities are needed.

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

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

Incidence:

- *13.2% as definition*
- *20.4% as interpretation*

[1] Jeffreys (1961).

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

Incidence:

- *13.2% as definition*
- *20.4% as interpretation*

Possible explanations:

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

[1] Jeffreys (1961).

4. The Bayes factors in applied research

Bayes factors are (at least can be) *sensitive* to priors

Very well known.[1,2,3,4,5]

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

Very well known.[1,2,3,4,5]

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).

This behavior of Bayes factors is in sharp contrast with estimation of posterior distributions.

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

This behavior of Bayes factors is in sharp contrast with estimation of posterior distributions.

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).

Reporting nothing at all (29.9%) or relying on software defaults (35.3%) was quite common.

Reporting nothing at all (29.9%) or relying on software defaults (35.3%) was quite common.

Possible explanations:

- Lack of awareness.
- Economic writing style.
- Default priors to...
  ...ease comparison, avoid specification, meet 'objectivity'.
  Also: improve peer-review chances, principle of indifference, preregistration.

4. The Bayes factors in applied research

Bayes factors are a measure of *relative* evidence

Say that $BF_{01} = 100$; what does this mean?

The observed data are 100 times more likely under $\mathcal{H}_0$ than under this particular $\mathcal{H}_1$.

[1] Morey, Romeijn, and Rouder (2016).
[2] Rouder (2014).
[3] Gelman and Rubin (1995).
[4] Ly, Verhagen, and Wagenmakers (2016).

Say that $BF_{01} = 100$; what does this mean?

The observed data are 100 times more likely under $\mathcal{H}_0$ than under this particular $\mathcal{H}_1$.

- Evidence is *relative*.[1]
- A model may actually be dreadful, but simply less so than its competitor.[2,3]
- Little is known as to how Bayes factors behave under model misspecification (but see[4]).

[1] Morey, Romeijn, and Rouder (2016).
[2] Rouder (2014).
[3] Gelman and Rubin (1995).
[4] Ly, Verhagen, and Wagenmakers (2016).

"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)."

Incidence: 62.3%

> *"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)."*

Incidence: 62.3%

Possible explanations:

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

4. The Bayes factors in applied research

Bayes factors can *not* establish absence/presence

Say that $BF_{01} = 100$, for $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \neq 0$.

This does not imply that $\mu = 0$.

Say that $BF_{01} = 100$, for $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \neq 0$.

This does not imply that $\mu = 0$.

- First of all, the Bayes factor (as the p-value) is a stochastic endeavor, not a factual proof.
- Furthermore, the Bayes factor provides a relative assessment of the likelihood of the observed data, not of the entertained hypotheses.

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

Incidence: 35.3%

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

Incidence: 35.3%

Possible explanations:

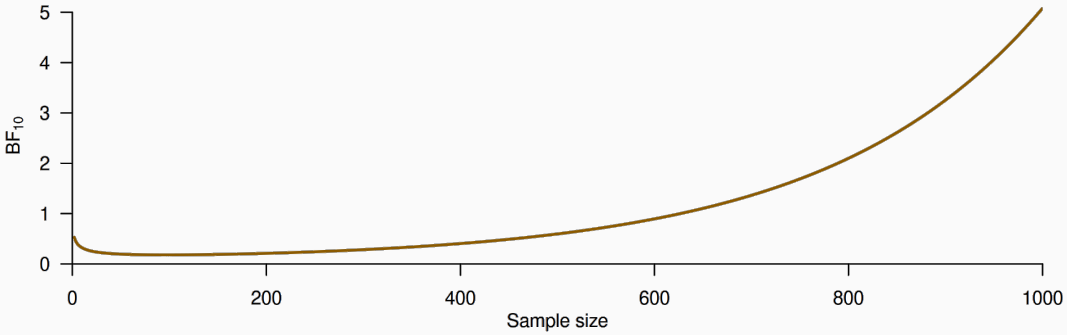
- Increased impact.
- Avoid uncertainty.
- Writing style.
- Influence from NHST.
- Decision making.

4. The Bayes factors in applied research

Bayes factors are *not* an effect size measure

Example:

- Bayesian one sample t-test:
  $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \neq 0$.
- JZS default prior ($r = .707$).
- $\overline{x} = 0.1$, $sd = 1$ at each sample size (thus, the effect size is fixed throughout).



"Pupil size was larger in a higher tracking load (...). However, the Bayesian test showed only positive, but smaller, effect of Load on tracking pupil size ($BF_{incl.} = 7.506$)."

Incidence: 4.2%

> *"Pupil size was larger in a higher tracking load (...). However, the Bayesian test showed only positive, but smaller, effect of Load on tracking pupil size ($BF_{incl.} = 7.506$)."*

Incidence: 4.2%

Possible explanations:

- Recreating a similar misconception based on p-values.
- Bayes factor labels in use.

4. The Bayes factors in applied research

Inconclusive evidence is *not* evidence of absence

$$BF_{01} = \frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)} = 1$$

Data are equally likely under either model.

$$BF_{01} = \frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)} = 1$$

Data are equally likely under either model.

Data are perfectly uninformative.

This does not equate to "*there is nothing to be found*".

"In contrast there was no difference in meaning between the thinking without examples and planning conditions; the Bayes factor provided anecdotal evidence in favor of the null ($BF_{10} = .86$)."

Incidence: 3.6%

> *"In contrast there was no difference in meaning between the thinking without examples and planning conditions; the Bayes factor provided anecdotal evidence in favor of the null ($BF_{10} = .86$)."*

Incidence: 3.6%

Possible explanations:

- Recreating a similar misconception based on p-values.
- Absence as default.
- Dichotomization.
- Increased impact.
- Preference for parsimony.

4. The Bayes factors in applied research

Bayes factors are a *continuous* measure of relative evidence

Bayes factors are a continuous measure of evidence in $[0, \infty)$.
For instance, if $BF_{01} > 1$ then

- The observed 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 $\mathcal{H}_1$.

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

Bayes factors are a continuous measure of evidence in $[0, \infty)$.
For instance, if $BF_{01} > 1$ then

- The observed 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 $\mathcal{H}_1$.

Q: Can "*more likely than*" be qualified?
A: Several categorizations of strength of evidence (what is weak?, moderate?, strong?) exist.[1,2,3,4]

But this is problematic in various ways.

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

"(...) In terms of Bayes factor (BF), evidence for greater disgust in the experimental group was strong ($BF_{10} > 10$), but there was only weak evidence for a difference in other emotions (BF_{10}'s < 3)."

Incidence: 5.4%

> *"(...) In terms of Bayes factor (BF), evidence for greater disgust in the experimental group was strong ($BF_{10} > 10$), but there was only weak evidence for a difference in other emotions (BF_{10}'s < 3)."*

Incidence: 5.4%

Possible explanations:

- Summary.
- Seeking authority.
- Avoiding criticism.
- Borrowing from the literature and JASP.
- NHST ('significant', 'not significant').

5. Conclusions, next steps

I think that, concerning testing:

- Model comparison (including hypothesis testing) is really important.
- However, and clearly, researchers test *way* too much.
- Testing says very little about how well a model fits to data.

And what about estimation?

I think that:

- Testing need not be a prerequisite for estimation, unlike what some advocate.[1]
- Estimation quantifies uncertainty in ways that Bayes factors simply can not.
- Estimating effect sizes (direction, magnitude) is crucial. Bayes factors ignore this!
- Avoiding the dichotomous reasoning subjacent to Bayes factors can help.

[1] Wagenmakers et al. (2018).

And what about estimation?

I think that:

- Testing need not be a prerequisite for estimation, unlike what some advocate.[1]
- Estimation quantifies uncertainty in ways that Bayes factors simply can not.
- Estimating effect sizes (direction, magnitude) is crucial. Bayes factors ignore this!
- Avoiding the dichotomous reasoning subjacent to Bayes factors can help.

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).

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.

Questions?