Robustness of null hypothesis Bayesian testing under optional stopping

Jorge N. Tendeiro

IMPS 2020 – July 14

University of Groningen

Joint work with Henk Kiers and Don van Ravenzwaaij.

AKA:
Sequential testing.

Definition:
Continuously testing a null hypothesis ($\mathcal{H}_0$) as data are collected until $\mathcal{H}_0$ is rejected.

Procedure:

1. Collect some data.
2. Perform the test (α and n_{max} chosen in advance):
   Compute p and...
   - ...if $p < \alpha$: STOP and retain $\mathcal{H}_1$.
   - ...if $p > \alpha$: Back to 1.
3. Continue until either conclusive evidence is found or n_{max} is reached.

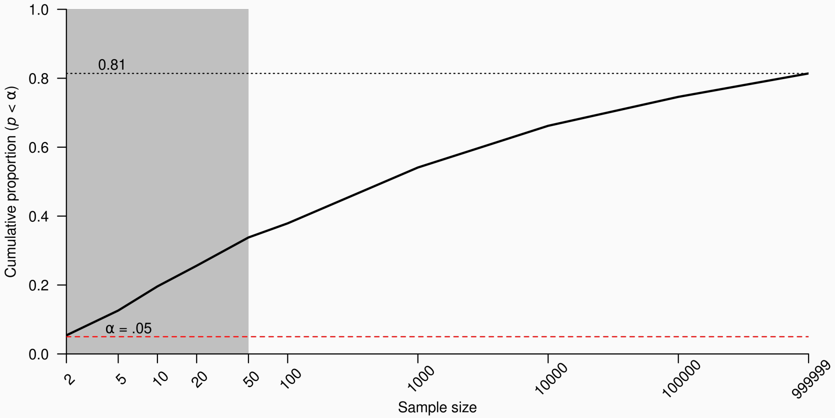
Known for a long time to be *very* problematic:

- Based on null hypothesis significance testing (NHST).
- NHST has a lot of problems.[1]
- In particular[2,3] : Too high proportions of false positives ($\gg \alpha$).

[1] Wasserstein, Schirm, and Lazar (2019). [2] Armitage, McPherson, and Rowe (1969). [3] Jennison and Turnbull (1990).

Example:

- One-sample t-test: $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_0 : \mu \neq 0$.
- Repeat 1,000 times:
  - Sampling plan: $n = 2(1)1,000,000$ from $\mathcal{N}(0,1)$.
  - Stop if $p < \alpha = .05$.



Some ways to avoid this problem:

- Using corrections[1,2,3,4,5,6,7,8] .
  Not commonly used in psychology.
- Not using optional stopping
  (i.e., fixed sample size, sample until completion).
- Turning to the Bayesian paradigm.

[1] Armitage (1960).
[2] Botella et al. (2006).
[3] Fitts (2010).
[4] Frick (1998).
[5] Jennison and Turnbull (1999).
[6] Lakens (2014).
[7] Pocock (1983).
[8] Wald (1945).

NHBT[1,2,3,4] is the Bayesian counterpart to NHST.

It uses the Bayes factor in place of the p-value.

Definition 1:
The Bayes factor quantifies the *update* in our relative belief about the likelihood of two hypotheses ($\mathcal{H}_0$, $\mathcal{H}_1$) in light of the observed data (D):

$$\underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{prior odds}} \times \underbrace{\frac{p(D|\mathcal{H}_1)}{p(D|\mathcal{H}_0)}}_{BF_{10}} = \underbrace{\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_0|D)}}_{\text{posterior odds}}$$

Definition 2:
The Bayes factor indicates the relative predictive value of each model.

- E.g., if the observed data are better predicted under $\mathcal{H}_1$ than under $\mathcal{H}_0$ then $p(D|\mathcal{H}_1) > p(D|\mathcal{H}_0)$ and so $BF_{10} > 1$.

[1] Jeffreys (1961).
[2] Kass and Raftery (1995).
[3] Tendeiro and Kiers (2019).
[4] van de Schoot et al. (2017).

Procedure:[1,2,3]

1. Collect some data.
2. Perform the test (BF_L, BF_U, and n_{max} chosen in advance):
   Compute BF_{10} and...
   - ...if $BF_{10} < BF_L$: Stop and retain $\mathcal{H}_0$.
   - ...if $BF_{10} > BF_U$: Stop and retain $\mathcal{H}_1$.
   - ...if $BF_L < BF_{10} < BF_U$: Back to 1.
3. Continue until either conclusive evidence is found or n_{max} is reached.

One *major* improvement of Bayesian over frequentist optional stopping:

> *The Bayesian procedure can stop due to sufficiently strong evidence in favor of $\mathcal{H}_0$.*

[1] Lindley (1957). [2] Edwards, Lindman, and Savage (1963). [3] Kass and Raftery (1995).

- It has been argued through the years that optional stopping under the Bayesian paradigm is allowed.[1,2,3,4,5]
- It has even been further develped and used in practice.[6,7,8,9,10]
- However, two recent papers disputed this state of affairs[11,12] (also[13]).

Rouder offered a rebuttal to these ideas in 2014.
Title: ‘Optional stopping: No problem for Bayesians’.[14]

---

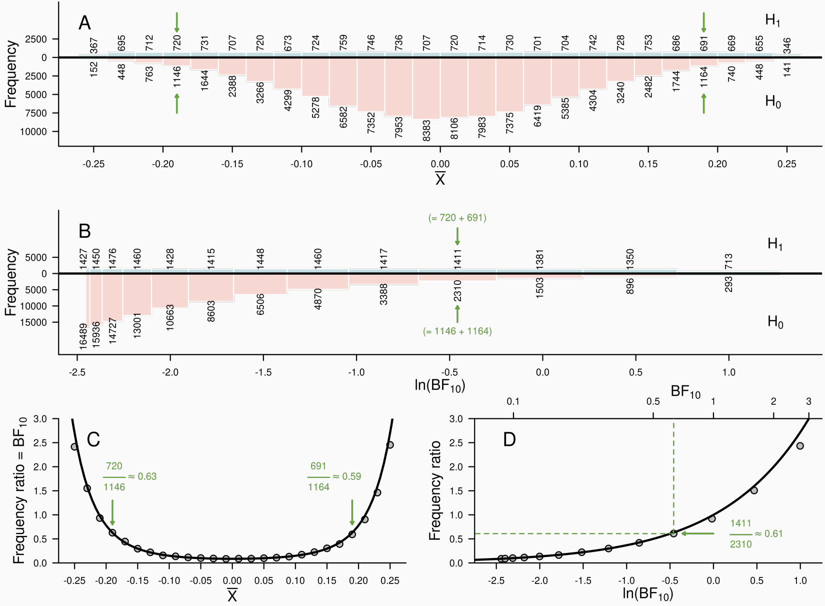
[1] Edwards, Lindman, and Savage (1963).
[2] Kass and Raftery (1995).
[3] Wagenmakers (2007).
[4] Wagenmakers et al. (2010).
[5] Francis (2012).
[6] Matzke et al. (2015).
[7] Schönbrodt et al. (2017).
[8] Schönbrodt and Wagenmakers (2018).
[9] Wagenmakers et al. (2012).
[10] Wagenmakers et al. (2015).
[11] Yu et al. (2014).
[12] Sanborn and Hills (2014).
[13] de Heide and Grünwald (2017).
[14] Rouder (2014).

Yu et al. (2014) and Sanborn & Hills (2014) questioned the *long run properties* of the Bayesian optional stopping procedure.

Rouder (2014) argued that there was no problem *in a particular sense*.

Let's visualize the argument:

- Data: $X_i \sim \mathcal{N}(\mu, \sigma^2)$, for $i = 1, \ldots, n$ and σ known.
- $\mathcal{H}_0 : \mu = 0$.
- $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, \sigma_1^2)$, for σ_1 known.



A
Frequency
H1
H0
367 695 712 720 731 707 720 673 724 759 746 736 707 720 714 730 701 704 742 728 753 686 691 669 655 346
152 448 763 1146 1644 2388 3266 4299 5278 6582 7352 7953 8383 8106 7983 7375 6419 5385 4304 3240 2482 1744 1164 740 448 141
2500 0 2500 5000 7500 10000
-0.25 -0.20 -0.15 -0.10 -0.05 0.00 0.05 0.10 0.15 0.20 0.25
X̄
B
Frequency
(= 720 + 691)
(= 1146 + 1164)
1427 1450 1476 1460 1428 1415 1448 1460 1417 1411 1381 1350 713
16489 15936 14727 13001 10663 8603 6506 4870 3388 2310 1503 896 293
5000 0 5000 10000 15000
-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0
ln(BF10)
C
Frequency ratio = BF10
720 / 1146 ≈ 0.63
691 / 1164 ≈ 0.59
0.0 0.5 1.0 1.5 2.0 2.5 3.0
-0.25 -0.20 -0.15 -0.10 -0.05 0.00 0.05 0.10 0.15 0.20 0.25
X̄
D
Frequency ratio
BF10
0.1 0.5 1 2 3
1411 / 2310 ≈ 0.61
0.0 0.5 1.0 1.5 2.0 2.5 3.0
-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0
ln(BF10)

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

The argument goes as follows:

- Assume prior odds equal to 1, so:

$$\underbrace{\frac{p(D|\mathcal{H}_1)}{p(D|\mathcal{H}_0)}}_{BF_{10}} = \underbrace{\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_0|D)}}_{\text{posterior odds}}.$$

- By definition of posterior odds:
  $\mathcal{H}_1$ is BF_{10} times more likely than $\mathcal{H}_0$ after considering the data.

Rouder made two assertions:

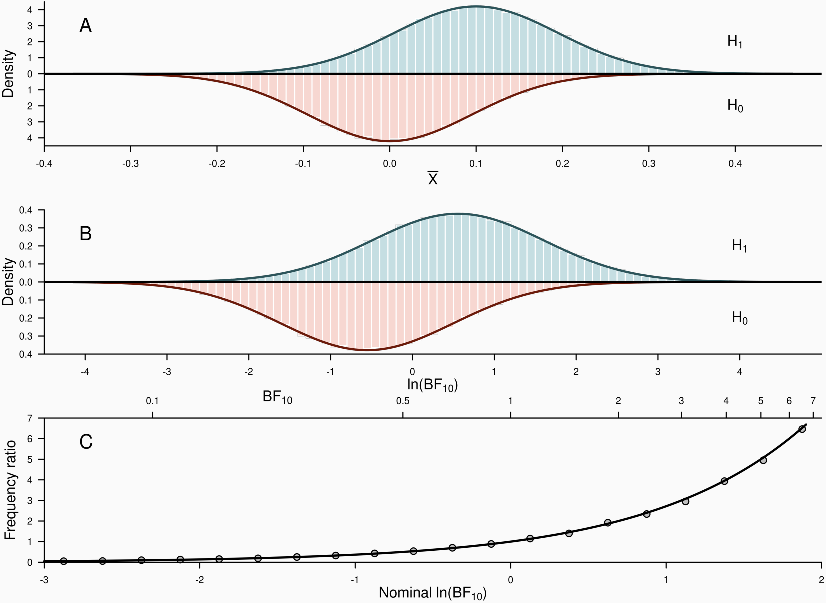
1. *Assertion 1:* For any given value BF_{10},
   $\mathcal{H}_1$ is BF_{10} times more likely than $\mathcal{H}_0$ to have generated BF_{10}.
2. *Assertion 2:* The above statement also holds under optional stopping.

In our paper, we:

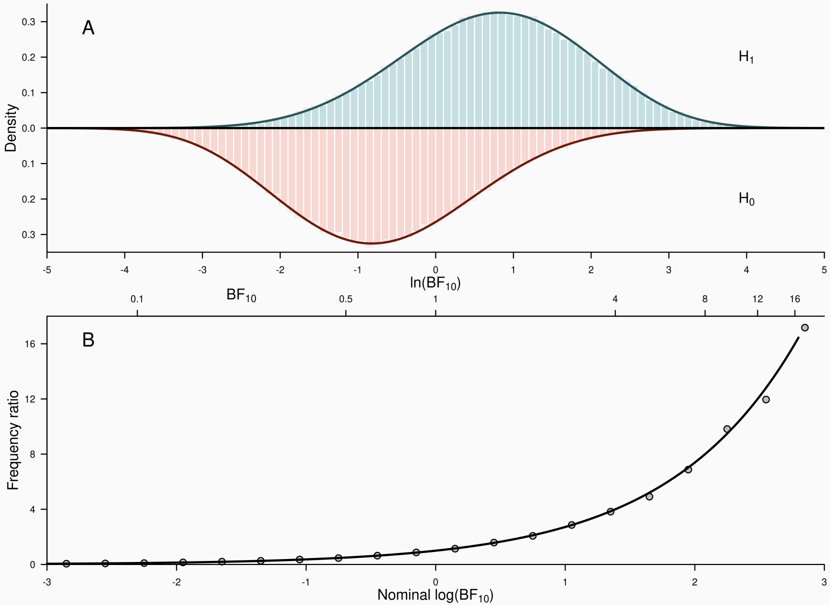
- Considered the same two tests as Rouder (2014):
  - Both tests about the mean of a normal distribution $\mathcal{N}(\mu, \sigma^2)$, σ known.
  - First test:
    $\mathcal{H}_0 : \mu = 0$ versus $\mathcal{H}_1 : \mu = \mu_1$.
  - Second test:
    $\mathcal{H}_0 : \mu = 0$ versus $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, \sigma_1^2)$, σ_1 known.
- Derived exact probability distributions for BF_{10}.
- Proved Assertion 1 for n fixed.
- Proved Assertion 2 after one step of the optional stopping procedure.

Results

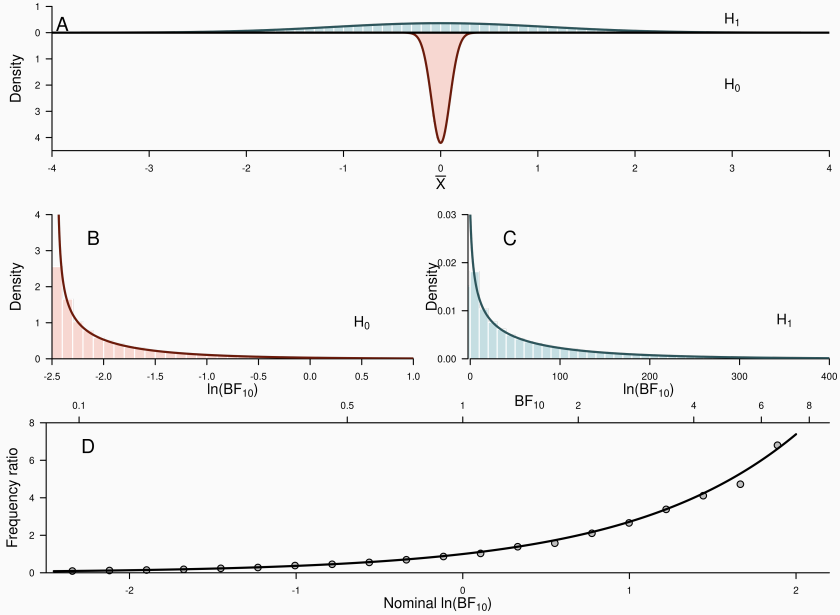
$\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = \mu_1$, n observations (Assertion 1)



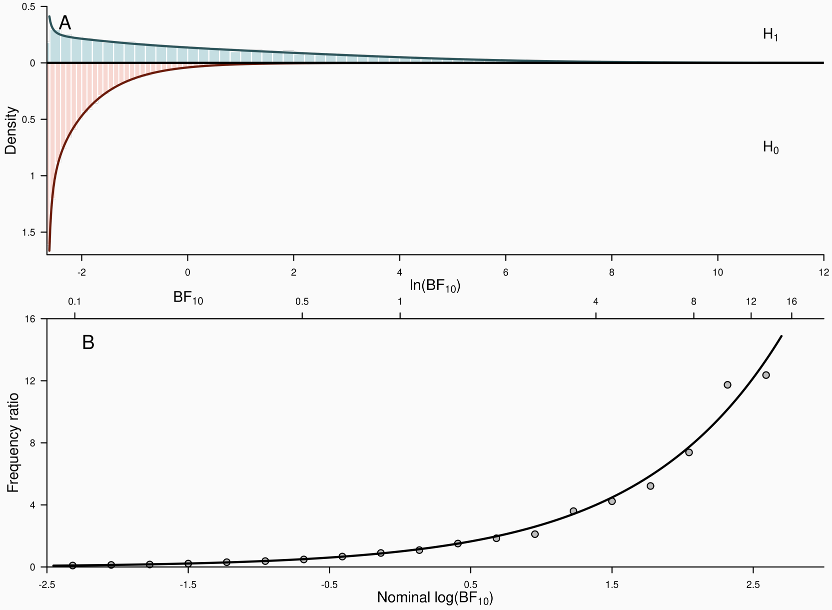
$\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = \mu_1$, $(n + k)$ observations (Assertion 2)



$\mathcal{H}_0 : \mu = 0$ VS $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, \sigma_1^2)$, n OBSERVATIONS (ASSERTION 1)



$\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, \sigma_1^2)$, $(n+k)$ observations (Assertion 2)



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] de Heide and Grünwald (2017).

Thank you!