bayesian sequential sampling in developmental science

peter zhu

04.05.24

some qualifications (and a roadmap)…

- I am not a statistician!
- A lot of this content is inspired by others' work
- Please feel free to interrupt with questions!
- I'm always happy to share materials or chat ☺ please reach out!

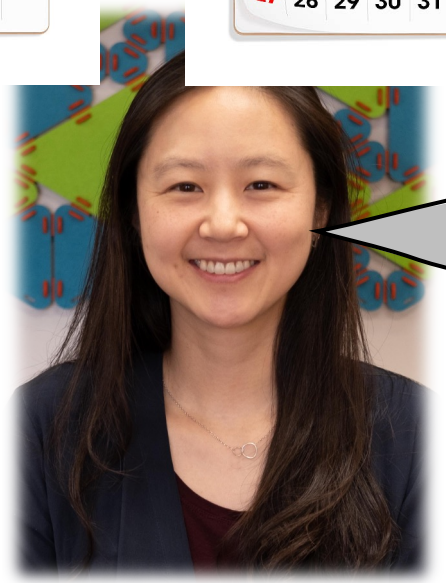
some qualifications (and a roadmap)…

Part 1:
An introduction to Sequential Bayes Factors (SBF)

Part 2:
How we've used SBF in our work

Part 3:
Pros + cons, and how to use it yourself

th y l developmental study...



Is there a difference in toy choice between toddlers in condition A and B?

Pilot, then preregister your study (and N)

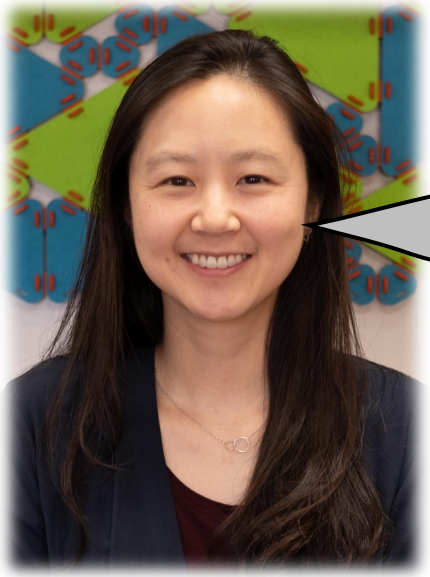
OSF

Data collection...





the life cycle of a developmental study...



Is there a difference in toy choice between toddlers in condition A and B?

Pilot, then preregister your study (and N)

OSF

Data collection...

Data analysis

Report your results

CogSci Proceedings: "We collected NN participants and found XX effect..."

the life cycle of a developmental study…

Is there a difference in toy choice between toddlers in condition A and B?

Some frustrations…

- Data collection is time-intensive, expensive, and hard!
- Assessing support for our hypothesis can be hard!
- What happens if the data are inconclusive? How do we know if there is no effect at all?

Pilot, then preregister your study (and N)

OSF

Data collection…

Data analysis

Report your results

CogSci Proceedings: "We collected NN participants and found XX effect…"

the life cycle of a developmental study…

Is there a difference in toy choice between toddlers in condition A and B?

Wouldn't it be nice if we could test as many participants as would give us confidence in an effect?

Pilot, then preregister your study (and N)

OSF

Data collection…

Data analysis

Report your results

CogSci Proceedings: "We collected NN participants and found XX effect…"

the life cycle of a developmental study…

Is there a difference in toy choice between toddlers in condition A and B?

One possibility: Sequential Bayes Factors (SBF)

- Also called Bayesian Sequential Sampling, with optional stopping
- In theory, this approach can reduce data collection, and increase confidence in results
- What's it all about?!

Pilot, then preregister your study (and N)

OSF

Data collection…

Data analysis

Report your results

CogSci Proceedings: "We collected NN participants and found XX effect…"

Oakes (2017)

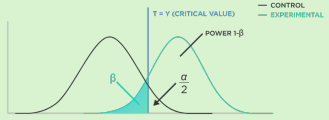
analytic strategy: classic (nhst)

This classic approach doesn't seem to solve our frustrations!

1. Preregister methods and N

Based on e.g., power analyses, prior work, or just picking an *n* that seems reasonable!

OSF



AS PREDICTED

How do we know what the right N is? (**We are often unsure!**)

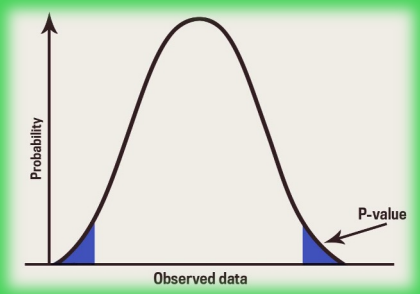
2. Collect data

Collect data according to your pre-registered sample size

We have to collect up to specified N, no more or less!

3. Analyze your data

Run appropriate statistical model or test; often use *p-value* to denote statistical significance



Is p < .05?

Interpretability issues! What happens if data is unclear?

For detailed approach: Schonbrodt & Wagenmakers (2018); Mani et al., (2020); Schonbrodt et al. (2017)

sbf as an alternative framework

Sequential Bayes Factors (SBF)

This piece helps with data collection!

This piece helps with interpreting data and drawing conclusions!

sbf as an alternative framework

Sequential Bayes Factors (SBF)

This piece helps with data collection!

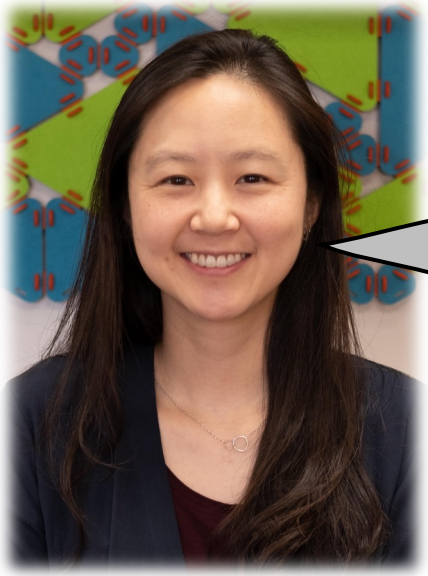
This piece helps with interpreting data and drawing conclusions!

the bf in sbf

posterior odds

$$\frac{P(H_1|D)}{P(H_0|D)}$$

Instead of using p-values, use Bayes Rule to calculate the *Posterior Odds*



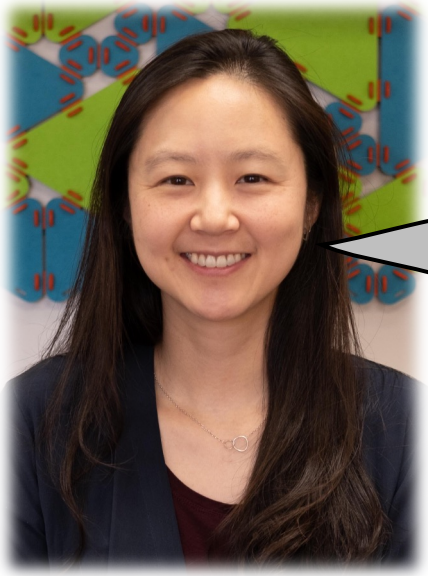
H_1: Condition effect on toddlers' toy choice!
H_0: No effect of condition!

the bf in sbf

Think of the Posterior Odds as how much *more likely* one hypothesis is over another

posterior odds

$$\frac{P(H_1|D)}{P(H_0|D)} = 5$$

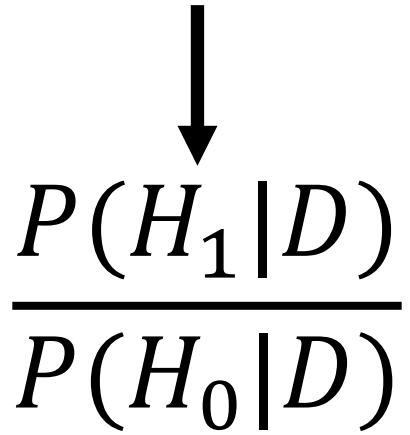


H_1: Condition effect on toddlers' toy choice!
H_0: No effect of condition!

H_1 is 5x more likely to be true than H_0

how does sbf work?

posterior odds



how does sbf work?

posterior odds

prior odds, which we often treat as 1 (equally likely)

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

Ratio of marginal likelihoods, or *Bayes Factor* (BF_{10})

how does sbf work?

posterior odds

prior odds, which we often treat as 1 (equally likely)

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

For our purposes, we treat the posterior odds the same as the **Bayes Factor**; the BF is what we will try to calculate!

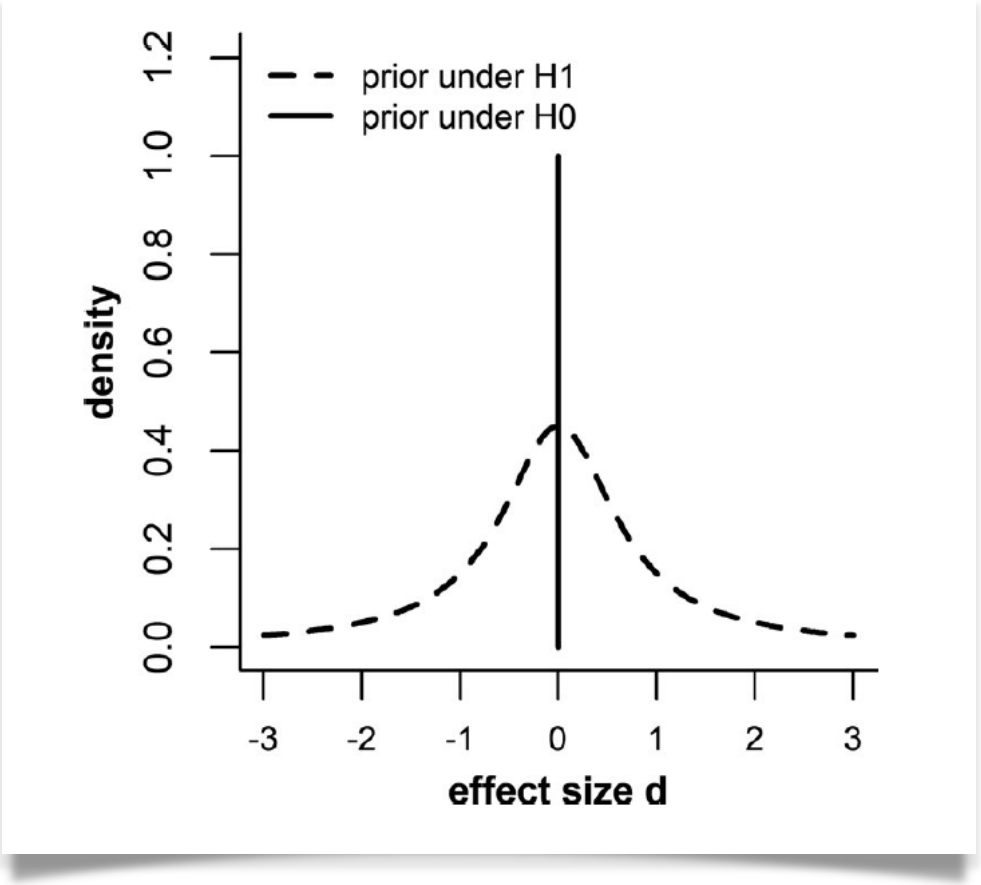
Ratio of marginal likelihoods, or *Bayes Factor* (BF_{10})

what's behind computing BF?

what is the probability of the data given a model?

Calculating the marginal likelihood involves defining a prior distribution e.g., over a given model parameter for each hypothesis

- Importantly, not the same as the prior odds



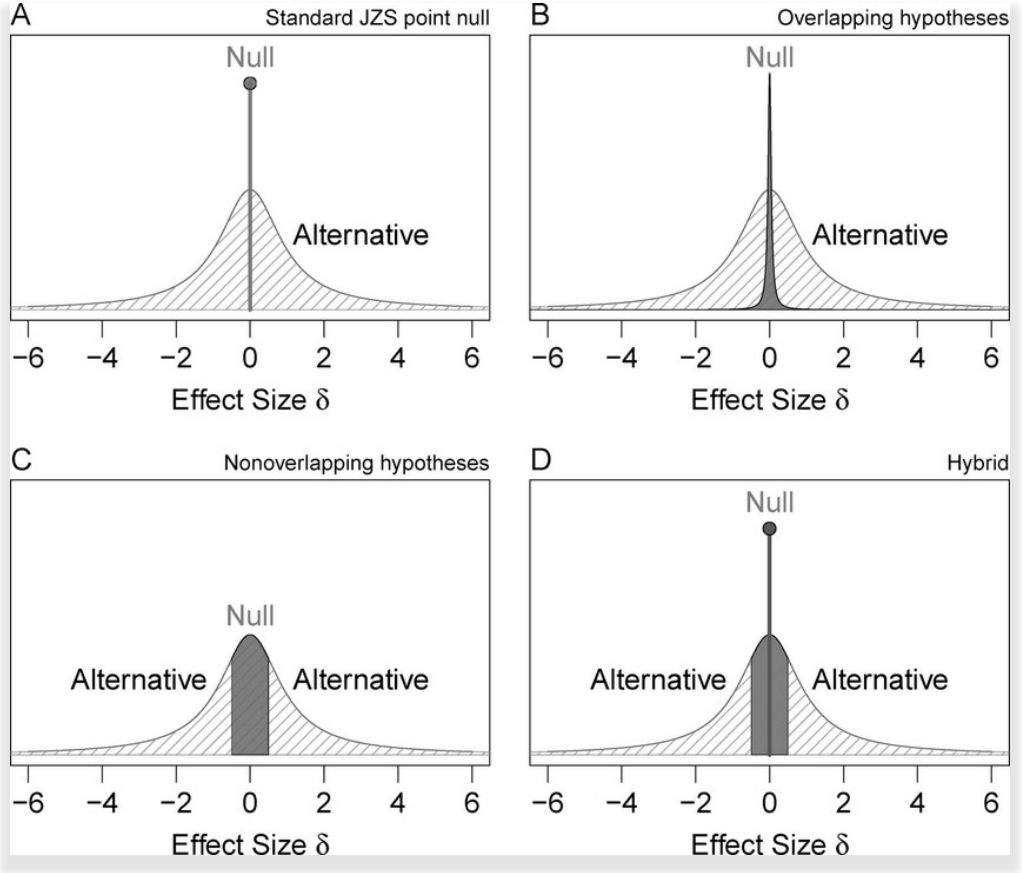
what's behind computing BF?

what is the probability of the data given a model?

Calculating the marginal likelihood involves defining a prior distribution e.g., over a given model parameter for each hypothesis

- Importantly, not the same as the prior odds

How you define this prior can have a huge impact on the Bayes Factor!



It can be tricky (and consequential) to try to define the priors yourself from scratch…

How to define good priors is the subject of intense debate in statistics and won't be the focus here… many resources are available[1]

Our recommendation: take an informed approach by looking at **default priors** implemented in various R packages (these folks thought a lot about this)!*

how do we compute the Bayes Factor?

two packages in R that make this easy!

1. The `BayesFactor` package

Build a Bayesian model object (e.g., `ttestBF`, `aovBF`, etc.)

Use the `summary` command to extract BFs

Lots of flexibility (and intuitive) to sample from the posterior, define various priors if you'd like, etc.

2. The `BFpack` package

Build a statistical model with usual R methods (e.g., with `stats: t.test`, `glm`, etc.)

BFpack takes these fitted model objects and performs BDA + hypothesis testing (using the command `BF`)

Calculates BFs slightly differently than the `BayesFactor` package!

putting the s in sbf

let's finally talk about the sequential part of SBF!

important terms:

Sequential Sampling

1. collect a pre-determined sample size *n*
2. calculate the BF
3. sample sequentially (in x units of participants or time) until you're confident in H_0 or H_1

Optional Stopping

Pre-determine thresholds (and/or max *n*) for interpreting evidence; if it passes threshold, you can *optionally* choose to stop collecting data

BF Value	Interpretation
> 30	Very Strong H1
10–30	Strong H1
3–10	Moderate H1
1–3	Anecdotal H1
1/3–1	Anecdotal H0
1/10–1/3	Moderate H0
1/30–1/10	Strong H0

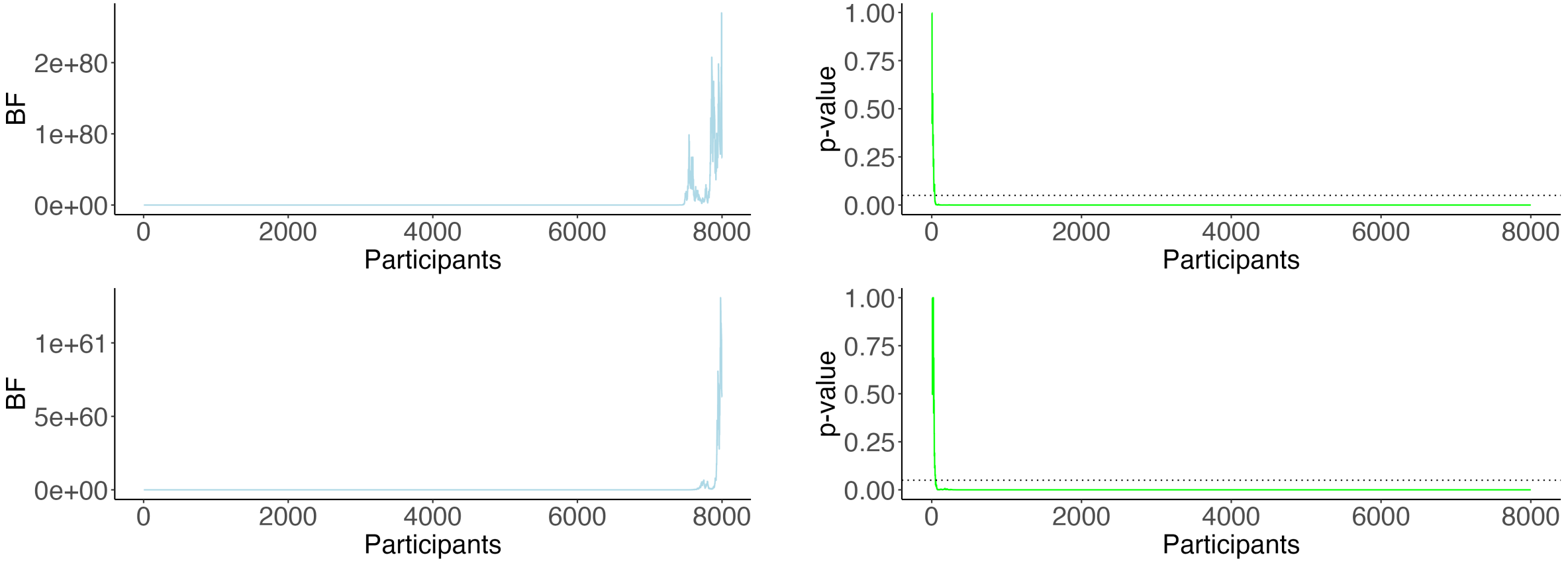
What does the sequential approach get us?

When an effect (null or predicted) effect is so clear and large (i.e., passes threshold), you can stop early and use resources to ask new questions!

- If you see evidence for H_1, you can begin working on follow-up studies!
- If you see evidence for H_0, you can be confident in abandoning a study and trying something else!

why not sequential with nhst?

If H_1 is true, then BF approaches ∞ and p–value approaches to 0

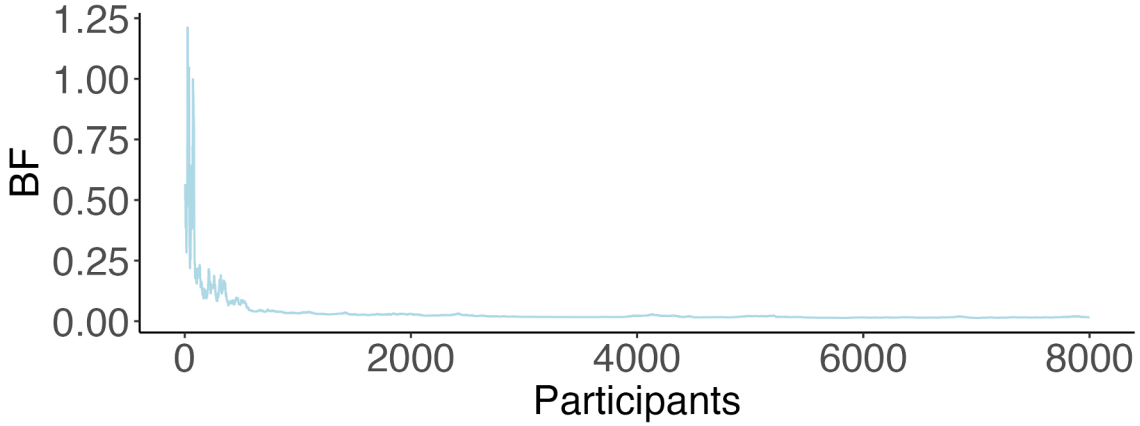


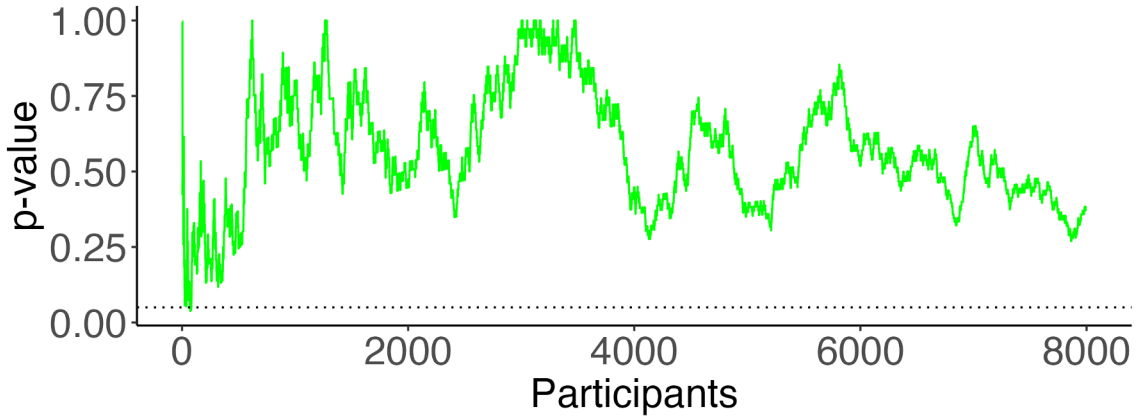
Rows: different random seeds

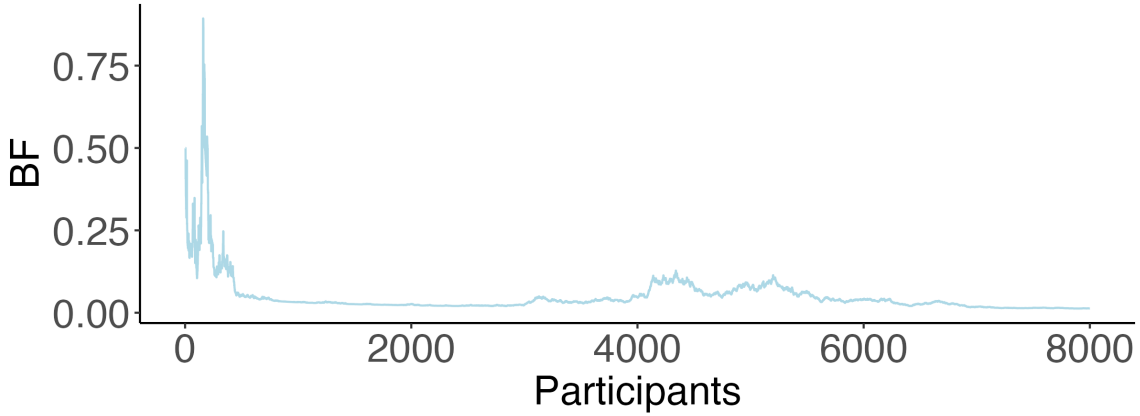
Analysis inspired by Schott et al., 2019

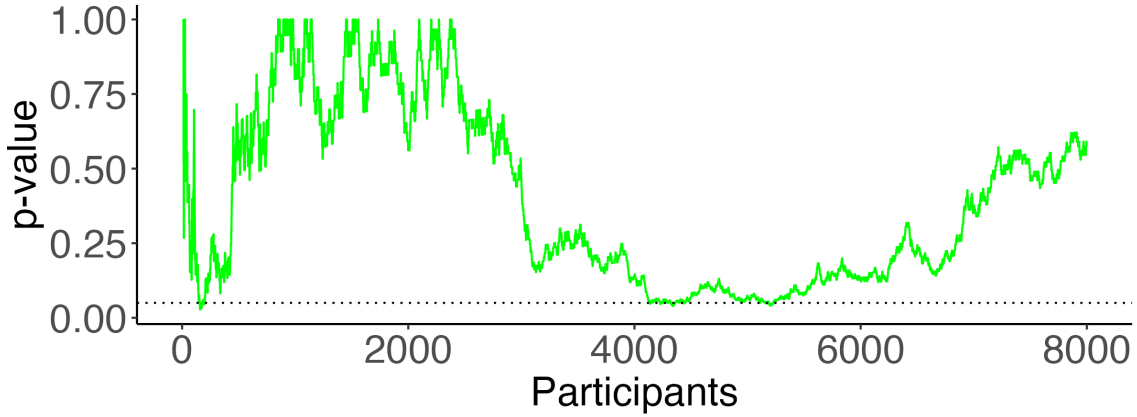
why not sequential with nhst?

However, if H_0 is true, BF approaches 0 while p-value shows no systematic pattern







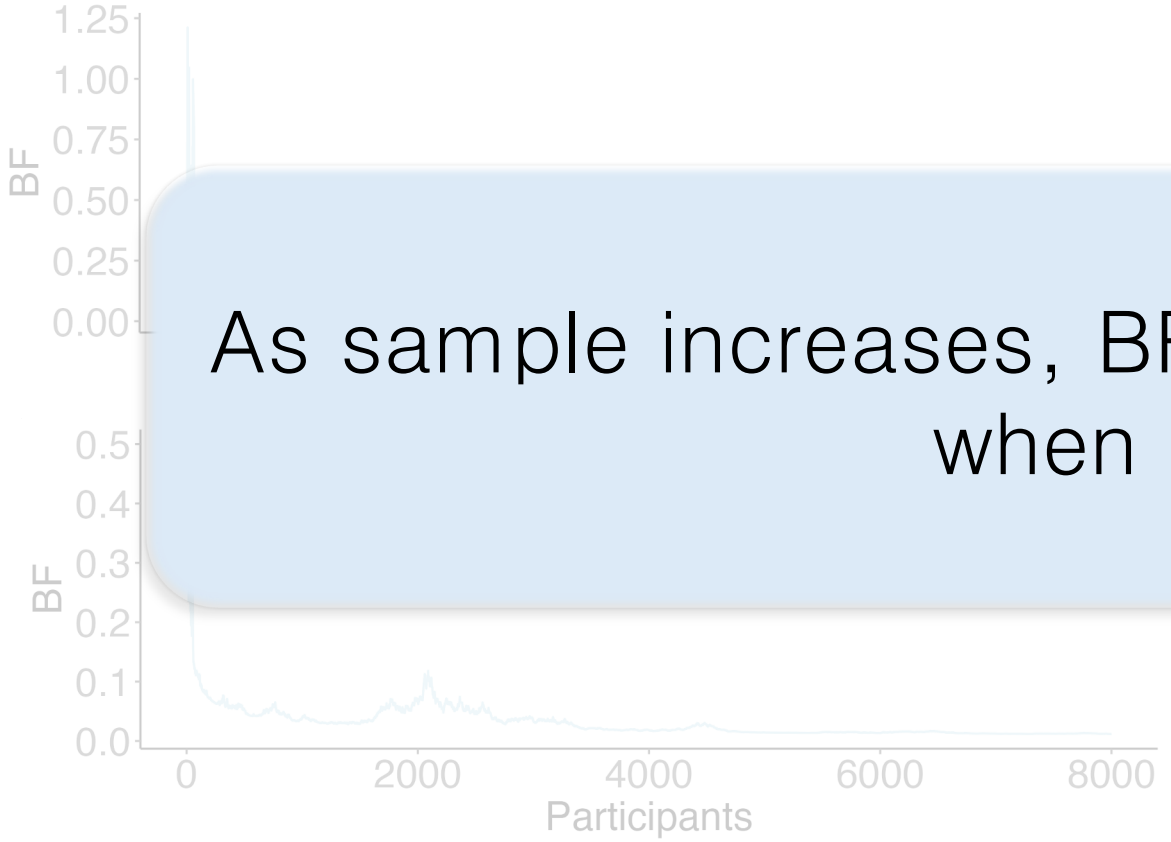


Rows: different random seeds

Analysis inspired by Schott et al., 2019

why not sequential with nhst?

However, if H_0 is true, BF approaches 0 while p-value shows no systematic pattern



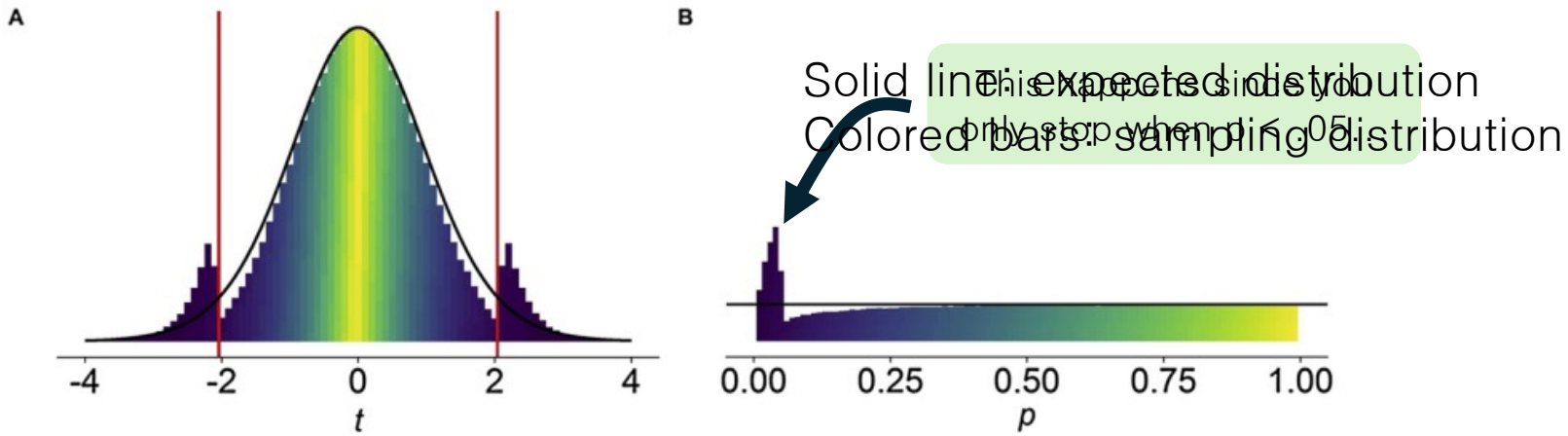
As sample increases, BF captures the effect even when it's null!

Schott et al., 2019

why not sequential with nhst?

Sequential sampling methods with NHST framework exist!

But repeatedly checking results with stopping rules increases Type-I error rates! Doing this analysis requires **careful correction of p-values**



Schott et al., 2019

why not sequential with nhst?

Sequential sampling methods with NHST framework exist!

But repeatedly checking results with stopping rules increases Type-I error rates! Doing this analysis requires **careful correction of p-values**

However, optional stopping with BFs do not inflate Type I errors in the same way… they quantify how compatible the available data is with either hypothesis!*

the sbf approach

1. Preregister sampling procedure

Start with ~20/cell, then specify the sequential sampling procedure

OSF

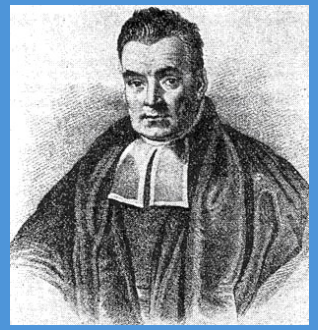
AS PREDICTED

2. Collect data

Collect data, evaluating BF sequentially until threshold or optional max N

3. Analyze your data

Calculate BF according to preferred model, can perform frequentist tests too!



What is the value of BF_{10}?

Don't need to explicitly worry about whether our N is "right"!

Only collect data until we are confident in an effect!

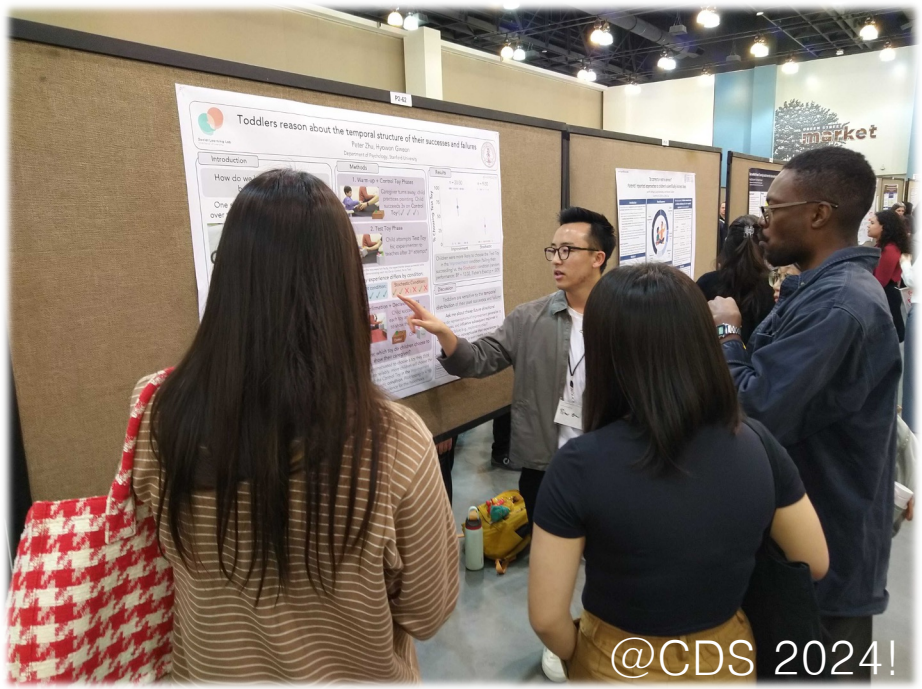
Data interpretation is straightforward, and only stop if we are convinced!

overview of sbf: summary

Four key takeaways…

1. BF approach uses Bayes Factors, rather than p-values, which quantify the probability of observing the data under a given hypothesis
2. This method allows us to do sequential sampling with optional stopping
3. It's not fundamentally different than your existing workflow! (probably)
4. We've been using the default priors already specified in various R packages (we like `BayesFactor` and `BFpack`)!

sbf in action: toddler study

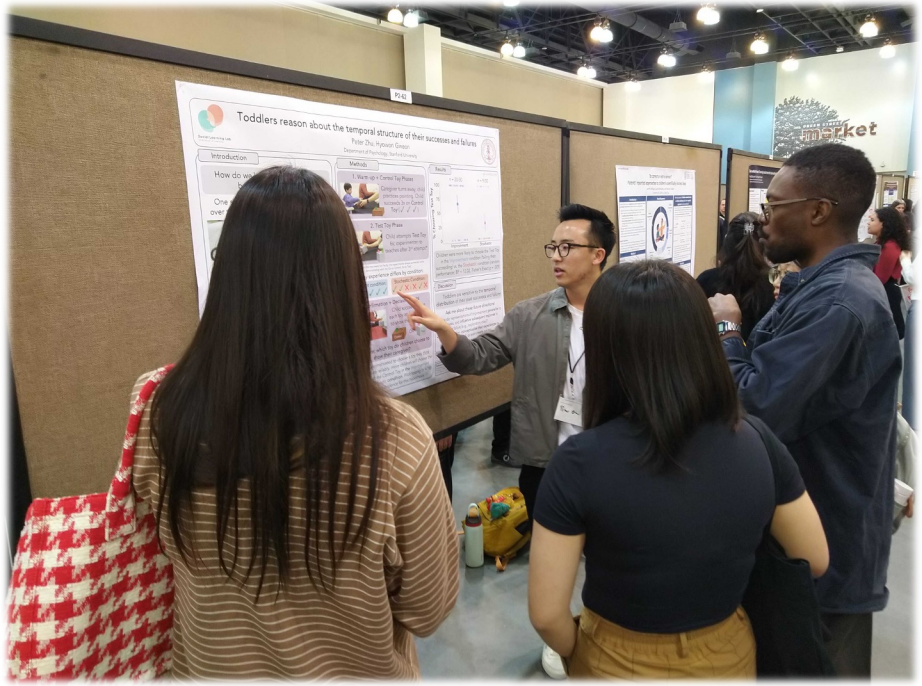


How do we know if we are getting better at something?

One straightforward way: tracking whether our **outcomes get better** over time

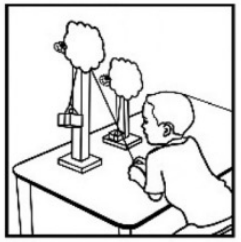
sbf in action: toddler study

Working with toddlers is hard… let's use SBF to minimize the pain!

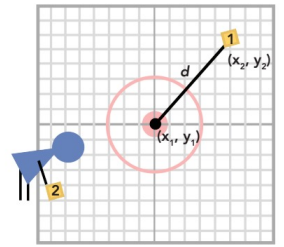


Children as young as 4 predict and track their performance over time…

But when does this ability emerge in life?



Leonard et al. (2023)

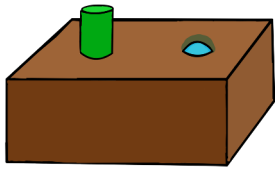


Zhang, Carrillo, & Leonard (2023)

Can two-year-olds represent temporal change in their performance over time?

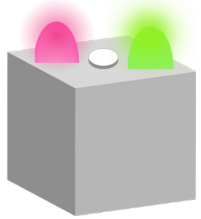
Caregiver is turned away, with headphones on

Control Toy Phase



Succeed 3x on the Control Toy

Test Toy Phase



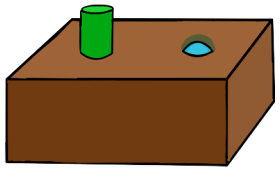
Children attempt Test Toy 6x; the experimenter “re-teaches” Test Toy after 3rd attempt

Caregiver is turned away, with headphones on

Control Toy Phase

Test Toy Phase

Decision Phase



Succeed 3x on the Control Toy

Activate each toy once more, then choose one to show parent

Improvement Condition:

✗ ✗ ✗ | ✓ ✓ ✓

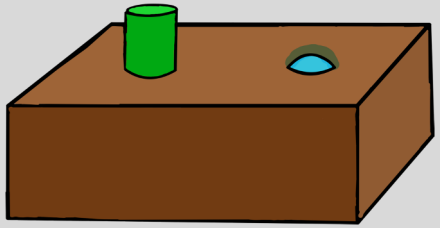
Stochastic Condition:

✓ ✓ ✗ | ✗ ✓ ✗

Key Dependent Measure

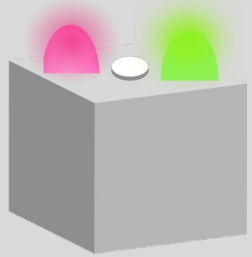
Children's choice during the decision phase: which toy do children choose to show their caregiver?

Control Toy



OR

Test Toy



If children are motivated to choose a toy they *think* they can activate reliably, more children should choose the Test Toy over the Control Toy in the Improvement than Stochastic condition (H_1: condition effect; H_0: null)

sbf in action: preregistration

The preregistration is *critical* for transparency with SBF methods; here's what we did!

Schonbrodt & Wagenmakers (2018)

sbf in action: preregistration

7) How many observations will be collected or what will determine sample size?

No need to justify decision, but be precise about exactly how the number will be determined.

We are uncertain as to the size of the effect we will observe (if any), thus we plan a sequential Bayes Factor analysis using the BFpack package in R (Mani et al., 2021). We will test an initial sample of 16 children (2-year-olds) in each condition, then evaluate the Bayes Factor on the hypothesis of a positive effect of condition on choice (i.e., Improvement condition -> higher likelihood of choosing the test toy) after 4 data points in each condition. We will stop testing at either a BF > 10 in favor of the hypothesis, a BF > 3 in favor of the null hypothesis (i.e., no difference in toy choice across conditions), or at an N = 36/condition (72 total).

Things I would do differently…

- Cite Mulder et al. (creators of BFpack) in addition to Mani et al. (SBF in developmental research)
- Run *n* = 20/condition for initial sample, rather than 16
- Specify how we would interpret the BF thresholds, and/or report development of BF
- Implement *optional* stopping, for flexibility to keep testing if we were unconfident in the effect

A quick note: notice the BF cutoffs for BF_{10} and BF_{01} are asymmetric; this is because evidence in favor of the null generally accumulates slower!

sbf in action: preregistration (what I would have done)

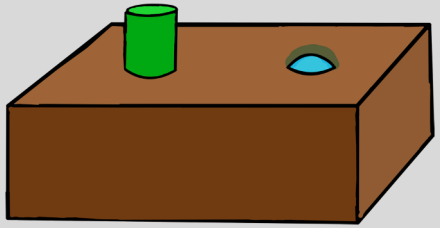
Sample Size
We are uncertain as to the size of the effect we will observe (if any), thus we plan a sequential Bayes Factor analysis using the `BFpack` package in R (Mulder et al., 2018; Mani et al., 2021). We will test an initial sample of 20 children (2-year-olds) in each condition, then evaluate the Bayes Factor on the hypothesis (H1) of a positive effect of condition on choice (i.e., Improvement condition -> higher likelihood of choosing the test toy) over no condition difference (H0) after 8 data points in each condition. We will stop testing at either a BF > 10 in favor of H1, a BF > 3 in favor of H0 (i.e., no difference in toy choice across conditions), or at an N = 36/condition (72 total).

Analysis (didn't mention in slides but probably good to change what we did; I would mention explicitly what the priors are)
We will conduct a 2x2 Contingency Table BF using the `contingencyTableBF` from the BayesFactor package in R. We will use the default prior as specified in the package (i.e., conjugate Dirichlet priors with $a = 1$; Gunel & Dickey 1987; Jamil et al. 2018)

Key Dependent Measure

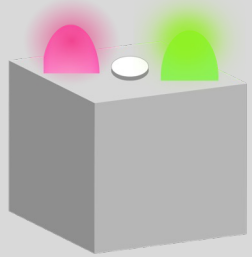
Children's choice during the decision phase: which toy do children choose to show their caregiver?

Control Toy



OR

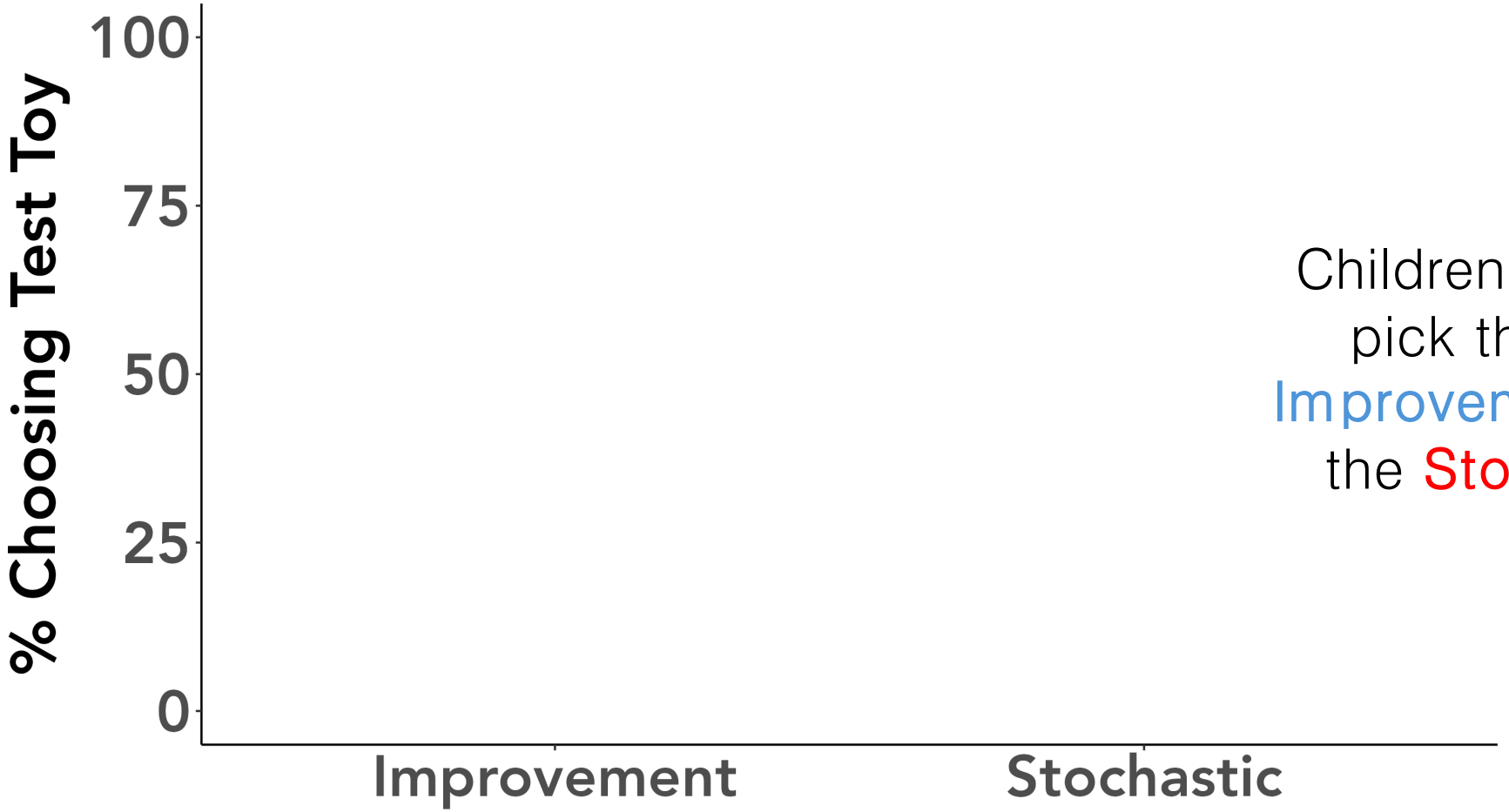
Test Toy



If children are motivated to choose a toy they *think* they can activate reliably, more children should choose the Test Toy over the Control Toy in the Improvement than Stochastic condition (H_1: condition effect; H_0: null)

sbf in action: results

Our SBF stopping rule led us to a total $n = 60$ (30/condition)



Children were more likely to pick the Test Toy in the Improvement condition, than the Stochastic condition!

$BF_{10} = 12.02$; Fisher's Exact $p = .009$

sbf in action: analysis

Let's switch to RStudio for a quick demonstration!

sbf in action: reporting results

Our recent *CogSci* submission:

Methods

fect of age, we limited our age range to two-year-olds. Also, given the relatively robust trend in Expt. 1, we preregistered a Bayesian sequential sampling analysis (Mani et al., 2021). This approach allows us to stop data collection in the presence of strong evidence in favor of the hypothesis (i.e., toddlers are more likely to pick the Test Toy in the Improvement condition; Bayes factor, or BF $>$ 10), weak evidence in favor of the null hypothesis (BF $>$ 3), or a final n of 72. [1]

sbf in action: reporting results

Our recent *CogSci* submission:

Results

Stochastic condition ($n = 9/28$, 32%). We conducted a logistic regression predicting toy choice from condition (assessing the data in support of the hypothesis of a condition effect over the null hypothesis of no difference across conditions). This analysis yielded a Bayes Factor (BF) of 8.05, which we interpret as moderate to strong evidence in favor of the predicted hypothesis. As an exploratory analysis, we conducted a Fisher's Exact Test in order to test the effect of condition on choice. This analysis revealed a significant effect of condition ($OR = 4.33$, 95% CI: [1.28, 15.95], $p = .015$).

sbf in action: reporting results

Our

Stoch
tic re
ing th
over t
This a
interp
dicted
a Fish
choice.
($OR = 4.33$, 95% CI: [1.28, 15.95], $p = .015$).

Things I'd do differently next time:

- More explicitly lay out what H1 and H0 are
- Be clear about the R packages used, what commands, and what the priors are!

But there are many ways to write results; no one "right" way!

sbf in action: reporting results

Really clear description of the sampling procedure and interpretation!

Kat's recent *CogSci* submission:

2021) with a contingency table BF, assessing the data supporting the hypothesis of a condition effect (H1) over the null hypothesis (H0: no difference between conditions). We interpret a $BF > 10$ as strong evidence, $BF > 5$ as moderate evidence, and $BF > 3$ as weak evidence for a condition effect. This analysis informed our sequential sampling procedure to determine the final sample size. We tested an initial sample of 10 children in each condition, and then evaluated the BF on the hypothesis of a condition effect after each day of testing. Stopping criteria was set at a $BF > 10$ in favor of the hypothesis of a condition effect (H1), a $BF > 3$ against the hypothesis (H0), or at an $n = 30$/condition ($N = 60$ total).

sbf in action: reporting results

Age distribution was not equal across conditions at first check; kept collecting data!

Kat's recent *CogSci* submission:

[1]Due to unequal distribution of age across conditions, we continued data collection despite reaching our sequential sampling stopping criterion of $BF(\text{H1}) = 19.17$ at $n = 28$. Age was better distributed across condition for the subsequent BF analysis at $n = 34$, and the criterion of $BF(\text{H1}) > 10$ was no longer met.

sbf in action: summary

Our results suggest that children as young as 2 years of age are sensitive to the *temporal distribution* of their past performance and use it to guide their future actions!

This representational capacity may be foundational to how children learn about themselves and their abilities!

Future Directions:

- How does this manifest in **exploratory play**?
- How do children **conceptualize their concrete experiences** as “success” or “failure”?
- When do children begin to represent **”graded” notions** of task performance?

sbf in action: summary

Our results suggest that children as young as 2 years of age are sensitive to the *temporal distribution* of their past performance and use it to guide their future actions!

This representational capacity may be foundational to how children learn about themselves and their abilities!

Future Directions:

- How does this manifest in **exploratory play**?
- How do children **conceptualize their concrete experiences** as "success" or "failure"?
- When do children begin to represent **"graded" notions** of task performance?

SBF was particularly useful here!

Using a sequential sampling method, we were able to stop data collection **12 points earlier** (!) than we would have otherwise!

A successful use of SBF!

the **pros** of sbf

1. Using SBF can reduce (in some cases, dramatically) the # of participants you test!
   - Allows the flexibility of testing additional children if you are unsure of the effect!

2. Using Bayesian analyses, you can quantify evidence for the ”null” hypothesis as well

3. Interpreting the statistical outputs is intuitive (at least more so than a p-value)

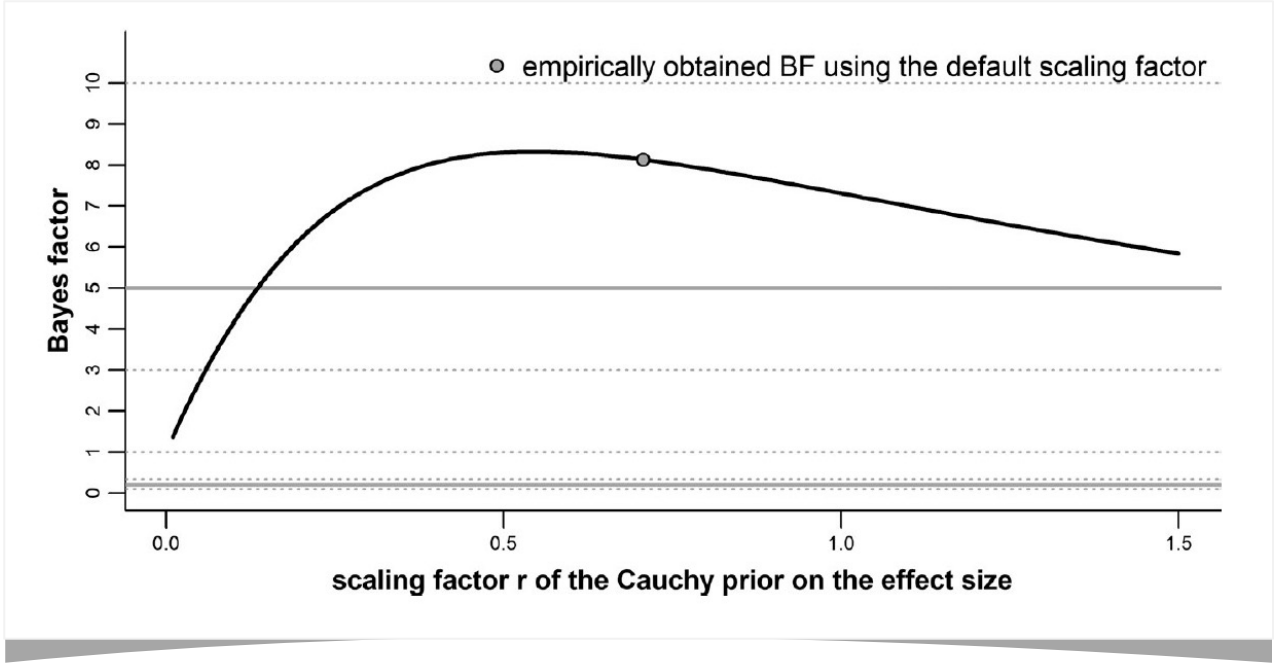
Mani et al., (2020)

the **cons** of sbf

1. Can be tricky in cases of relatively complex designs or statistical models
   - i.e., ones without common default priors (you will have to define these… the BF can be very volatile

One solution: conduct a variety of sensitivity analyses…

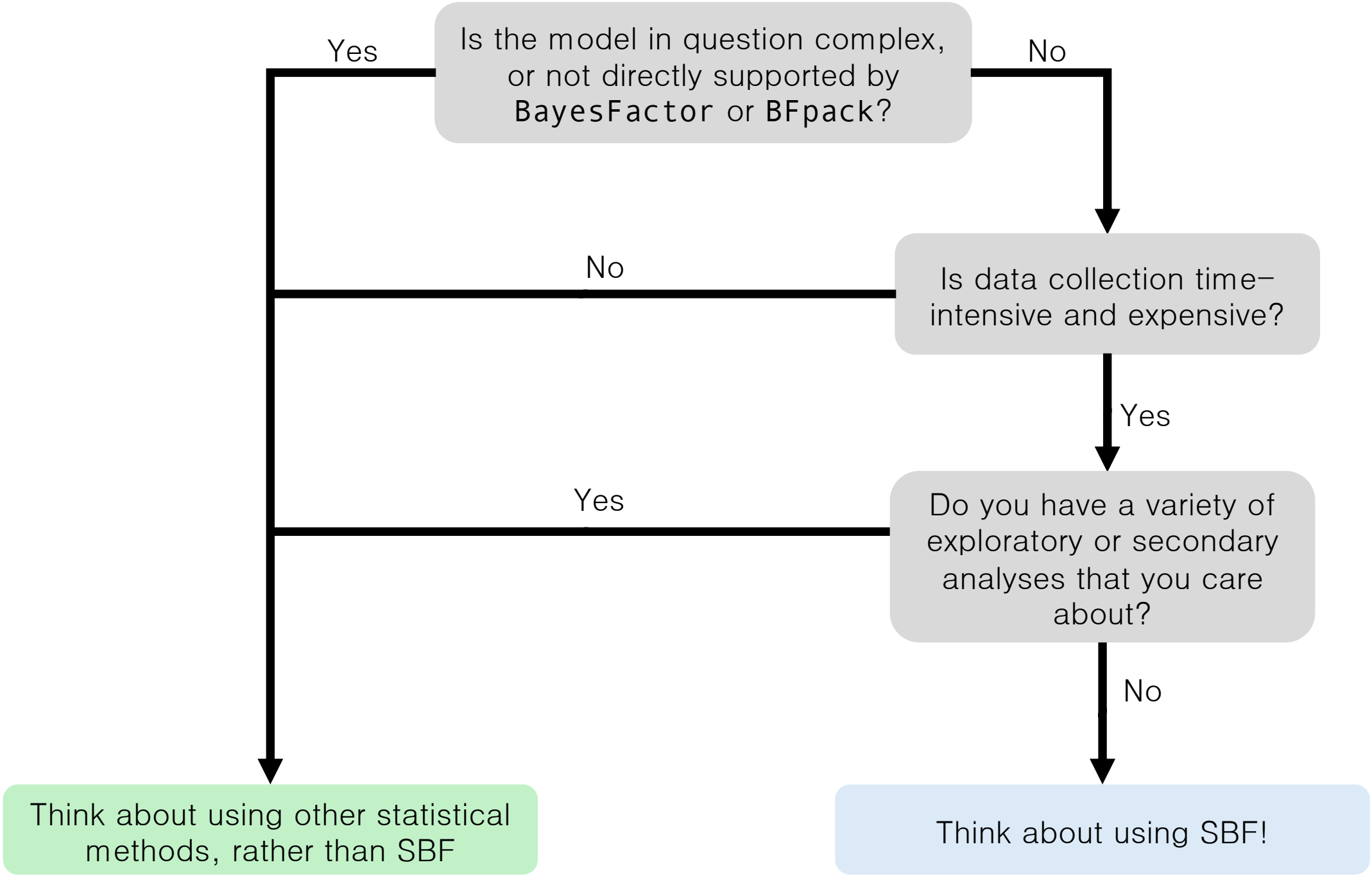
But doing so can be difficult and tricky!



the **cons** of sbf

1. Can be tricky in cases of relatively complex designs or statistical models
2. Gets a bit funky when running a variety of analyses (e.g., age effects)
3. Modifications needed when data collection is not costly or time-consuming (e.g., Lookit)
4. Is not perfect, and can still end up with inconclusive results if you hit your max *n*
5. Other concerns (i.e., data peeking is stressful!)

should you use sbf?


Is the model in question complex, or not directly supported by BayesFactor or BFpack?
Yes
No
Is data collection time-intensive and expensive?
No
Yes
Do you have a variety of exploratory or secondary analyses that you care about?
Yes
No
Think about using other statistical methods, rather than SBF
Think about using SBF!

conclusions

SBF can be a useful and powerful tool that can alleviate frustrations of developmental science

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

It can help alleviate the costs of testing, and aid with data interpretability… but it's not without its costs!

…maybe it could be useful to you as you run your own experiments!

thanks so much!

PALO ALTO junior MUSEUM & ZOO

BING NURSERY SCHOOL STANFORD

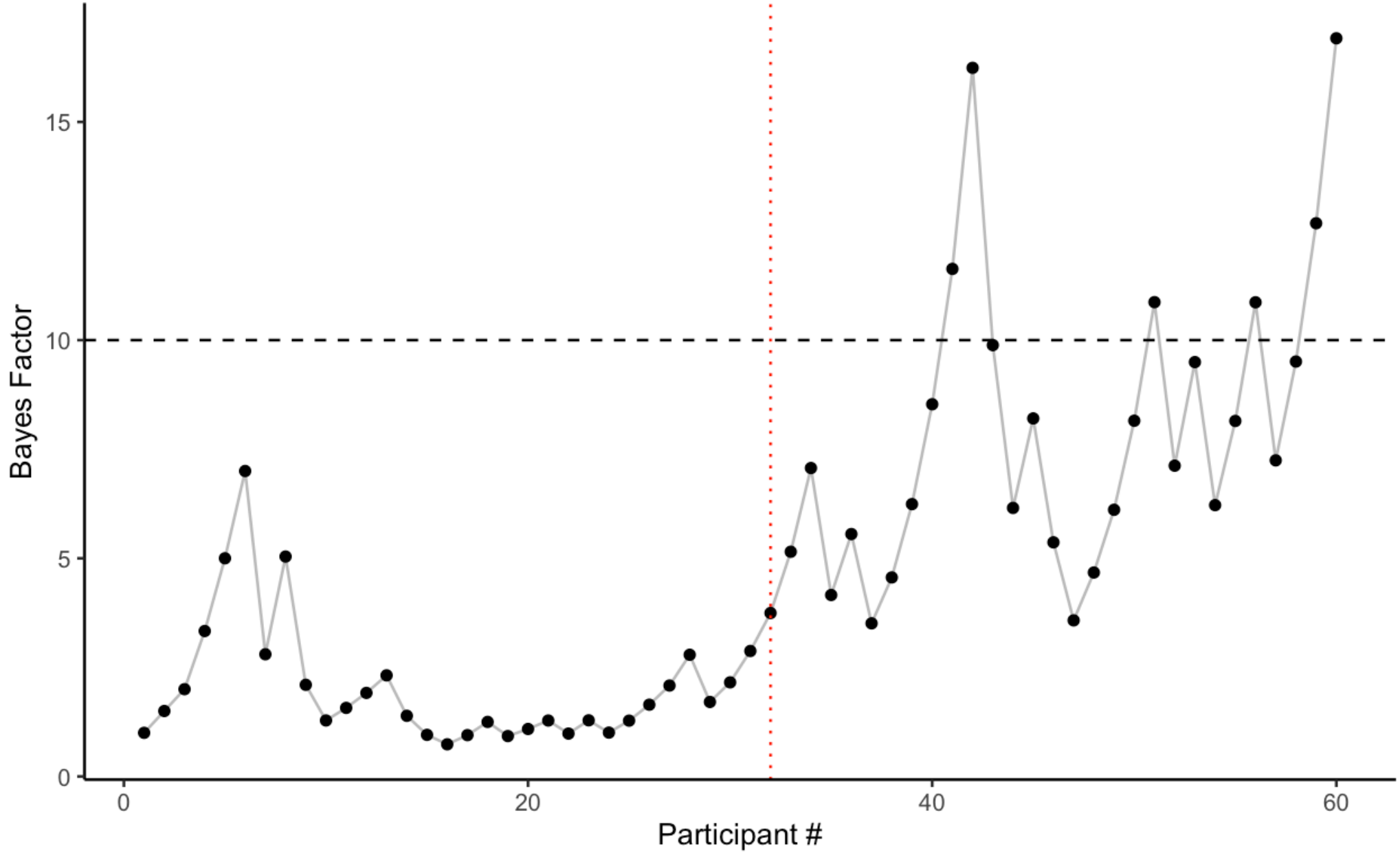
Hyowon Gweon

and also…

Alvin Tan
Mike Frank
Kat Adams Shannon
Emily Chen
Grace Keene
Veronica Aranda
Ellen Aasted
Libby Rouffy

Scan for link to repo with materials (slides, papers, code, etc.)!

time for a tutorial?


Bayes Factor
15
10
5
0
0
20
40
60
Participant #