Evaluating Evidence and Making Decisions using Bayesian Statistics

ISCoP Conference 2021

Mattan S. Ben-Shachar

tinyurl.com/ISCoP-2021-bayes

Presented at February 23, 2021

(updated February 23, 2021)

About Me

Mattan S. Ben-Shachar

PhD Student + Stats Lover + R Developer

Ben-Gurion University of the Negev
Beer Sheva, Israel

@mattansb | @mattansb

About You

👉 You use statistical models (in R)

- ANOVAs, regression
- Maybe some mixed models

👉 You've heard about (and maybe even used) Bayes factors

👉 **You want to know *more* about Bayesian stats**

🖵 Link to this presentation:
tinyurl.com/ISCoP-2021-bayes

All the code and materials used in this workshop can be found on GitHub:
github.com/mattansb/bayesian-evidence-iscop-2021

Outline

- What is a Bayesian model?
- How to Bayes, even?
- Why to Bayes? (aka "Why is this better than how I currently model?")
- Demo: Building a Bayesian model
    - Posterior Estimates
    - **Evaluating Evidence and Making Decisions using Bayesian Statistics**

Let us begin...

It's all About the ~~Bass~~ Bayesian Modeling

What *is* a Bayesian model?

A Bayesian model is a statistical model where you use **probability** to represent **all uncertainty** within the model, both the uncertainty regarding the output but also the uncertainty regarding the input (aka parameters) to the model[1]...

... where probability expresses *a degree of belief* in an event.

[1] Bååth (2015). *From stackexchange*

How to Bayes? 🧠

To fit a Bayesian model you need...

A Prior

A probability distribution representing your prior *belief* about the probability of possible values each parameter can take.

> *"Sounds too subjective to be used in Science!"*
>
> - You (2021)?

In real life applications, you would be hard-pressed to just use whatever prior you like - you would need to somehow **justify your prior** (which requires domain specific knowledge).

Similar to how you must also justify and use a reasonable likelihood function.

Watch also Bürkner (2018). *Why not to be afraid of priors (too much)*

A Likelihood Function

What process best describes the (conditional) data generation process?

For example:

- A binomial likelihood function for **binary** data
- A Poisson likelihood function for **count** data
- A cumulative multinomial likelihood function for **ordinal** data
- An inverse Gaussian / ex-Gaussian / [other] likelihood function for **reaction times**
- ...
- A Gaussian likelihood function for **conditionally normal** data

The likelihood function tells us the *probability of observing our data given the value(s) of some parameter(s)*.

This function is used to ***update the priors***, resulting in ***The Posterior***...

Prior + Likelihhod = Posterior

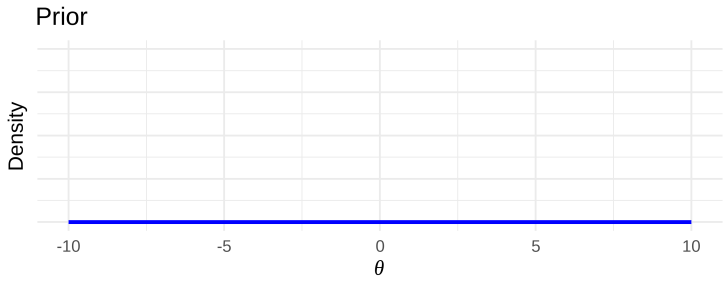
This is that whole pesky *Bayes' Rule* thing everyone keeps going on about:

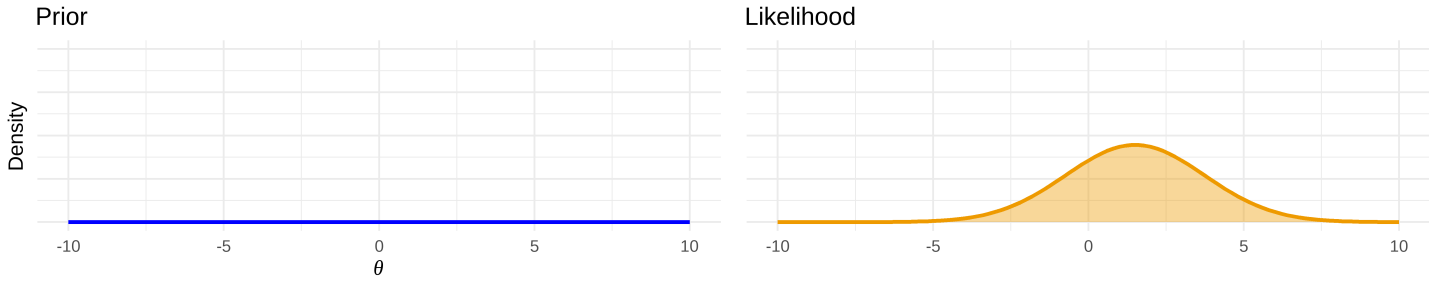
$$\overbrace{P(\theta|Data)}^{\text{Posterior}} = \frac{\overbrace{P(Data|\theta)}^{\text{Likelihood}} \times \overbrace{P(\theta)}^{\text{Prior}}}{P(Data)}$$

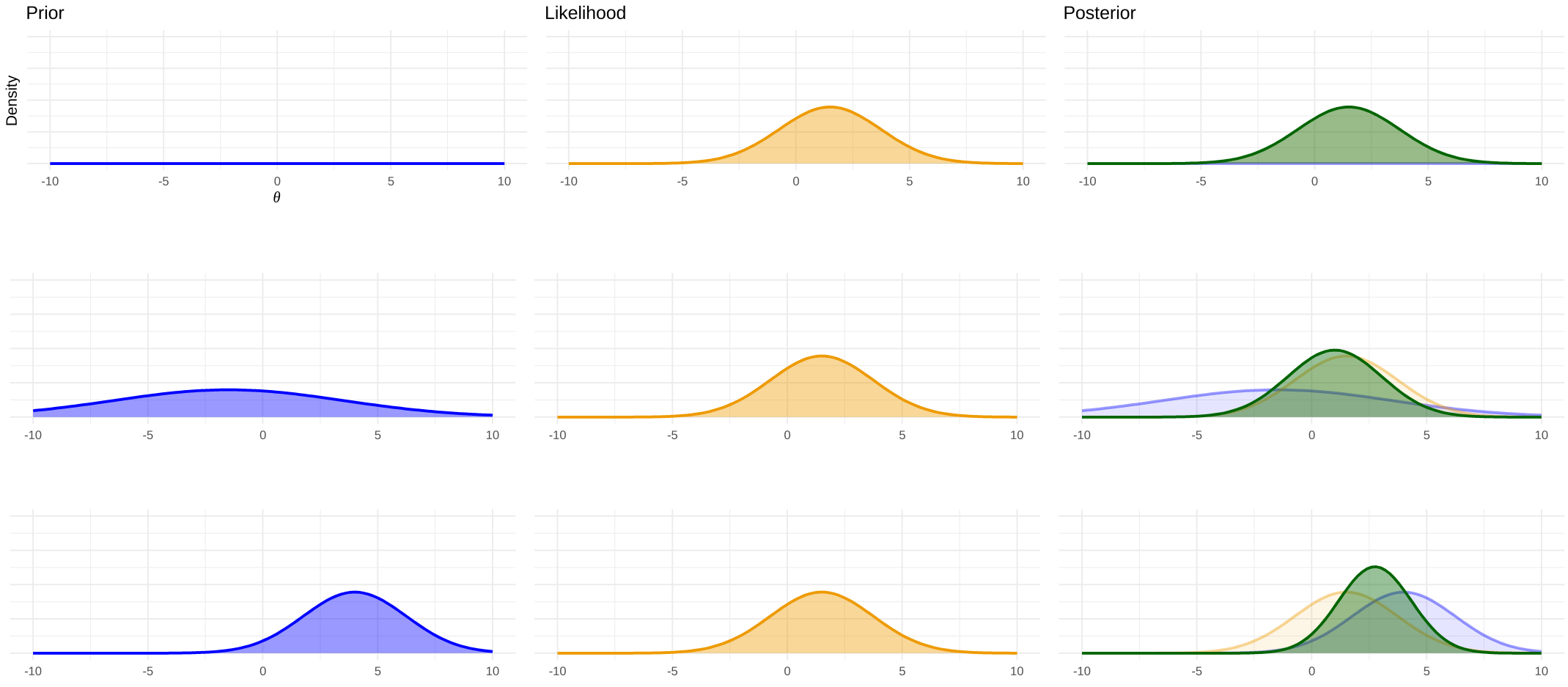
In words:

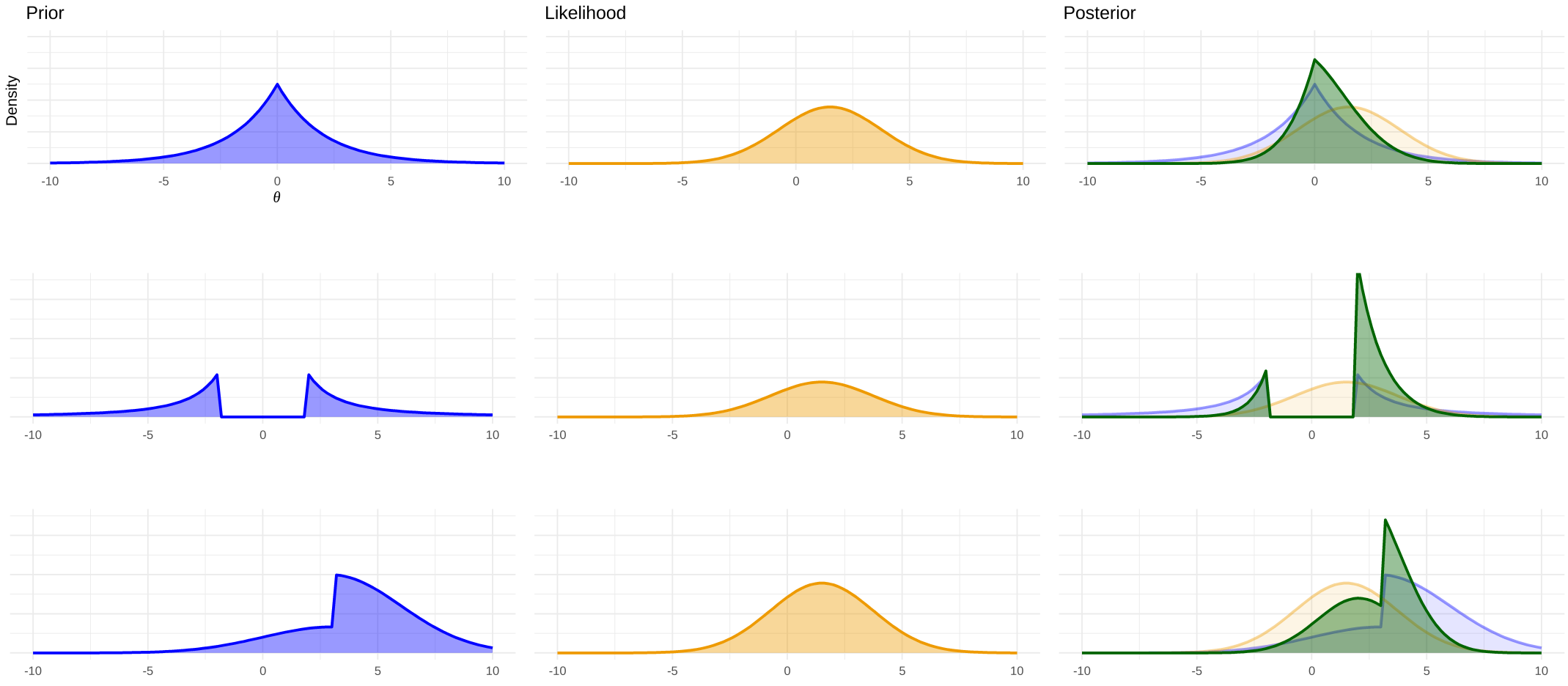
The **posterior probability** of some parameter θ having a value of x, is a equal to probability of the observed data occurring if that were the value of θ (**the likelihood**), normalized by our **prior belief** that θ can have a value of x.

We usually can only estimate the posterior distribution by sampling from it.

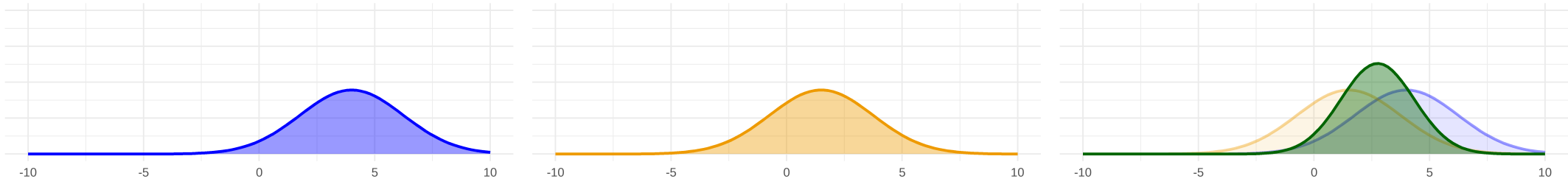

Prior
Density
-10
-5
0
5
10
θ


Prior
Likelihood
Density
-10
-5
0
5
10
θ

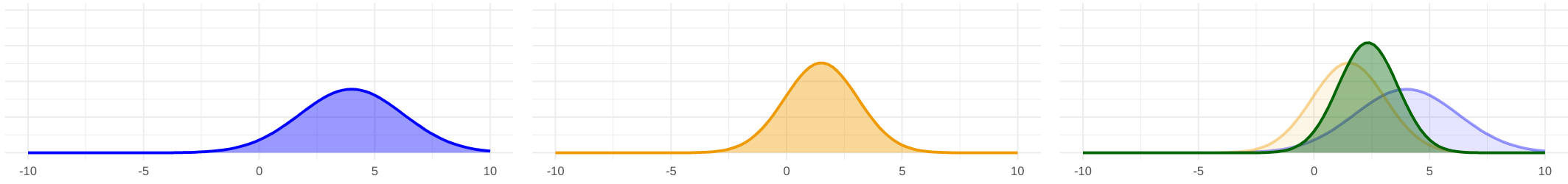

Prior
Likelihood
Posterior
Density
θ
-10
-5
0
5
10


Prior
Likelihood
Posterior
Density
θ
-10
-5
0
5
10

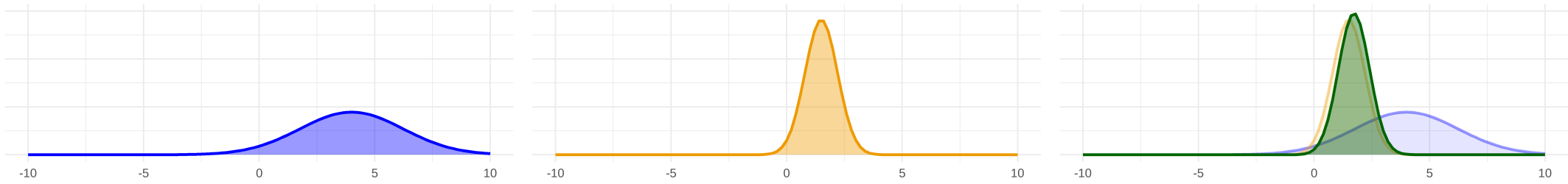
N = 30


-10
-5
0
5
10
-10
-5
0
5
10
-10
-5
0
5
10

N = 60


-10
-5
0
5
10
-10
-5
0
5
10
-10
-5
0
5
10

N = 300


-10
-5
0
5
10
-10
-5
0
5
10
-10
-5
0
5
10

Why to Bayes?

AKA "Why is this better than what I currently do?"

- **Speak in the language of probabilities** (*probabilitese?*).

  There is a 0.2 (posterior) probability of the treatment alleviating more than 3 ADHD symptoms.

  There is a 0.85 (posterior) probability of realibility of the test being at least $\alpha > 0.8$.

- **The power of Priors**
  - Utilize prior knowledge - *add* the information gained from the current data to the existing corpus of knowledge.
    - Not every study is *tabula rasa*.
  - Use priors to prevent over-fitting (regularization via horseshoe, spike-and-slab).

Why to Bayes?

AKA "Why is this better than what I currently do?"

Fit complex models / to complex data:

- Limiting the search space of our model's parameters to what is *a-priori* reasonable, reduces issues that plague other estimation methods.
    - failed convergence, local maxima, complete separation...
- With a likelihood function and a prior, you can add endless complexity to your model (even allow $n < p$).
    - Easily model heteroscedasticity,
    - Model individual differences in ICC in HLM,
    - Easily obtain CIs for random effects,
    - ...
- **Some types of models cannot practically be analyzed using frequentists methods** 🤯 (Rouder & Lu, 2005)

Demo

Let's get our hands dirty...

See the full analysis script here 🔗

We will be looking at a regression model,

but the tools from this demo can be applied to Bayesian SEM, IRT, SDT, etc...

The Data

Thirty 4 year old children completed **Flanker's task**. (real data.)

Congruent

Incongruent

Neutral

We will be examining their **Interference** (Incongruent - Neutral) and **Facilitation** (Neutral - Congruent) effects, **controlling for age** (in months).

Image from pixy.

We will be working in **R** with the following packages:

- `brms` for Bayesian Regression Models with *Stan*.
    - *Stan* is a probabilistic programming language
- `emmeans` for extracting estimates / contrasts / slopes from the model.
- `bayestestR` for descriptive and inferential statistics.
- Plots are made with `ggplot2` + `patchwork` + `tidybayes` + `ggdist` + `see`.

See other package versions and packages used 🔗

Building a Bayesian Model

```
m_flanker <- brm(
  RT ~ Congruency + age_mo + (Congruency | id),
  data = child_flanker,
  prior =
    # Two parameters for Congruency
    set_prior("student_t(3, 0, 100)", class = "b",
              coef = c("Congruency1", "Congruency2")) +
    # Slope of age_mo
    set_prior("student_t(3, 0, 1000)", class = "b",
              coef = "age_mo"),
  family = gaussian())
```

We will be fitting an hierarchical linear model - predicting (single trial) RTs from `Congruency` (I, N, C) which is nested within each child (`id`) - controlling for the children's age (in months, `age_mo`).

This is essentially a repeated measures ANCOVA.

Note: For `Congruency` I've used ***orthonormal*** **dummy-coding**. This is important, but ⏳! Read more about that here.

Building a Bayesian Model

```
m_flanker <- brm(
  RT ~ Congruency + age_mo + (Congruency | id),
  data = child_flanker,
  prior =
    # Two parameters for Congruency
    set_prior("student_t(3, 0, 100)", class = "b",
              coef = c("Congruency1", "Congruency2")) +
    # Slope of age_mo
    set_prior("student_t(3, 0, 1000)", class = "b",
              coef = "age_mo"),
  family = gaussian())
```

For our fixed effects, we will be somewhat conservative and use a scaled *t*(3)-prior centered on 0. This prior has the benefit of the scaling factor giving the range where 60% of the prior's mass is.

- In adults, the Flanker effect is about 20-50ms. Here we have 4yo - reasonable (?) that any differences between means (effect) would be **~100ms**, which we will use as our scaling factor (Jonkman et al, 1999).
- Prior on effect of age - no idea. We will use a weakly informative prior scaled to 1000ms/month (covering a very large range of possible effects).

Building a Bayesian Model

```
m_flanker <- brm(
  RT ~ Congruency + age_mo + (Congruency | id),
  data = child_flanker,
  prior =
    # Two parameters for Congruency
    set_prior("student_t(3, 0, 100)", class = "b",
              coef = c("Congruency1", "Congruency2")) +
    # Slope of age_mo
    set_prior("student_t(3, 0, 1000)", class = "b",
              coef = "age_mo"),
  family = gaussian())
```

Notes:

- By default, `brms` sets **flat** (*diffused, extremely uninformative*) priors for fixed effects.
- We can also set a prior of `sigma` (error variance), and many others. See more options with `brms::get_prior()`.

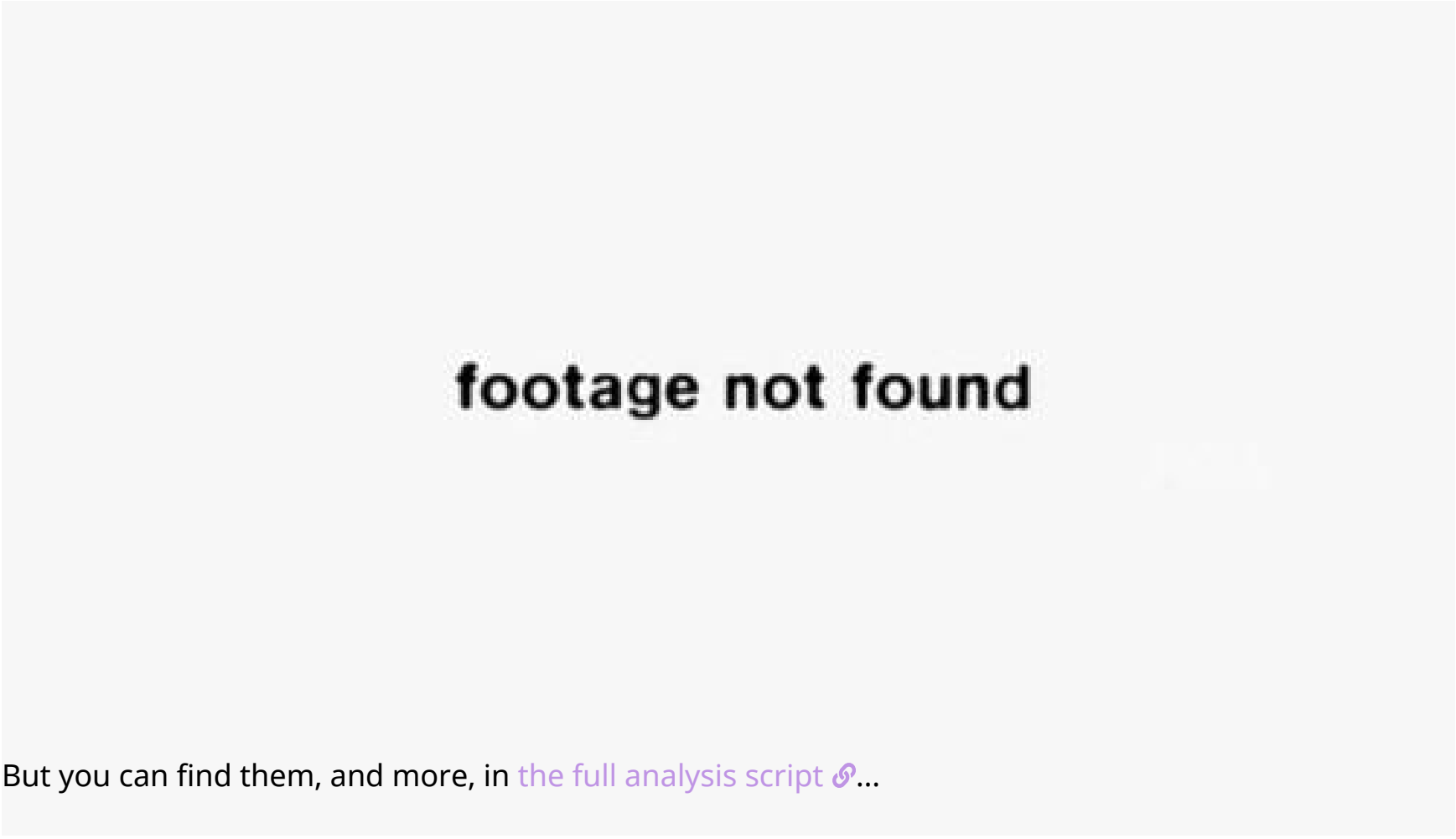
Building a Bayesian Model

```
m_flanker <- brm(
  RT ~ Congruency + age_mo + (Congruency | id),
  data = child_flanker,
  prior =
    # Two parameters for Congruency
    set_prior("student_t(3, 0, 100)", class = "b",
              coef = c("Congruency1", "Congruency2")) +
    # Slope of age_mo
    set_prior("student_t(3, 0, 1000)", class = "b",
              coef = "age_mo"),
  family = gaussian())
```

We will be using a Gaussian likelihood function of $RT \sim N(\mu_i, \sigma^2)$, where $\mu_i = a + \sum b_j X_{ij}$.

AKA, a boring linear regression.

Prior & Posterior Checks



But you can find them, and more, in the full analysis script 🔗...

Explore the Model

Let's look at the posteriors of the estimated means for the Congruency conditions:

```
means_Congruency <-
  emmeans(m_flanker, ~ Congruency)
```

Explore the Model

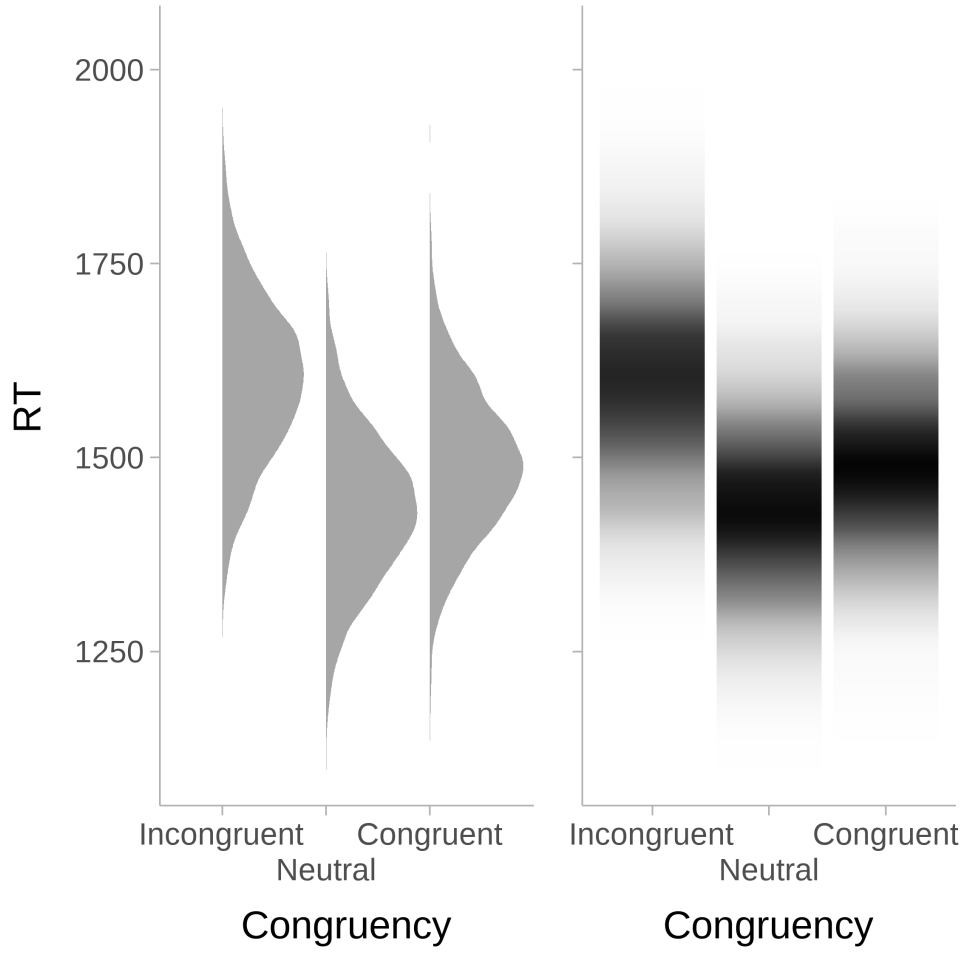
Let's look at the posteriors of the estimated means for the Congruency conditions:

```
means_Congruency <-
  emmeans(m_flanker, ~ Congruency)
```

Frequentist estimation methods (such as **OLS** or **maximum likelihood (ML)**) produce a point estimate for each parameter.

But in Bayes we have not a single value, but a whole distribution of values!

We we can either present the whole distribution, *as is*...



Or we can summarize the posterior distribution:

A Representative Value

(in lieu of a point estimate)

- Median (most common)
- Mean
- Maximum A Posteriori (MAP)

Credible Intervals (CIs)

- The Highest Density Interval (HDI; most common)
- The Equal-Tailed Interval (ETI)

```
describe_posterior(means_Congruency,
                   centrality = "median",
                   ci = 0.89, ci_method = "hdi",
                   test = NULL)
```

```
## # Description of Posterior Distributions
##
## Parameter   |   Median |              89% CI
## ------------------------------------------
## Incongruent | 1602.487 | [1420.115, 1753.119]
## Neutral     | 1433.930 | [1286.052, 1587.072]
## Congruent   | 1490.332 | [1341.278, 1639.822]
```

Evaluating Evidence and Making Decisions using Bayesian Statistics

We are limiting our discussion to evaluating evidence for **single estimates / parameters** (expected values, slopes, contrasts...).

But it is also possible to evaluating evidence for multiple parameters, with order restrictions and model comparisons. (Maybe next year...)

We will be looking at two contrasts: the Interference and Facilitation effects:

```
diffs_Congruency <- contrast(means_Congruency,
                             list(Interference = c(1, -1, 0),
                                  Facilitation = c(0,  1, -1)))

describe_posterior(diffs_Congruency, test = NULL)
```

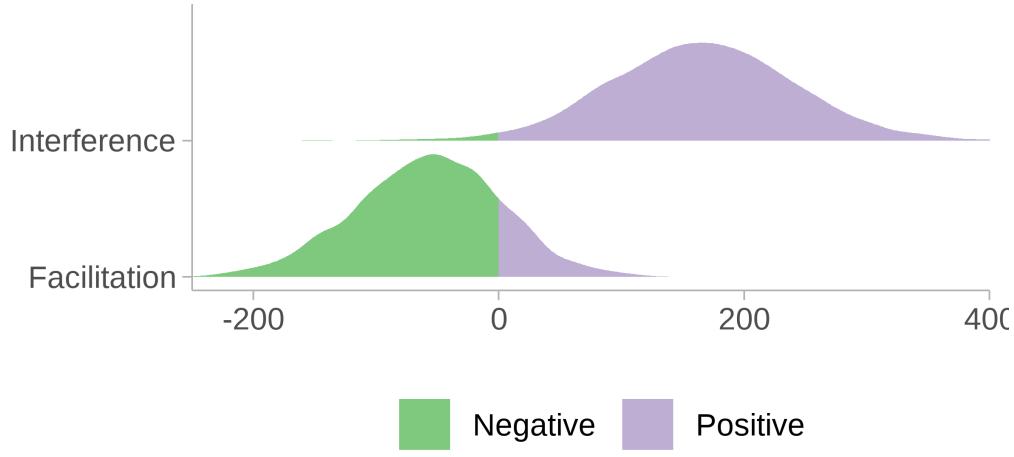
```
## # Description of Posterior Distributions
##
## Parameter    |  Median |              89% CI
## ----------------------------------------------
## Interference | 166.484 | [  48.509, 285.054]
## Facilitation | -56.012 | [-154.245,  35.098]
```

The Probability of Direction

- The maximal probability of the estimate being strictly directional (larger or smaller than 0).
- Generally ranges from 50% (no preference) to 100%.

```
p_direction(diffs_Congruency)
```

```
## # Probability of Direction (pd)
##
## Parameter    |     pd
## ---------------------
## Interference | 98.58%
## Facilitation | 83.58%
```



For the Interference effect it seems like these is a high probability of direction, but not that great for the Facilitation effect (p_d < 0.95).

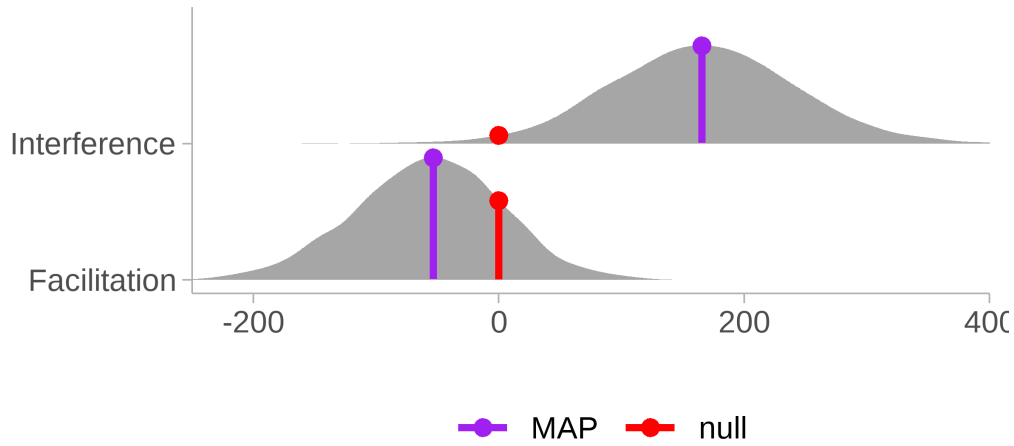
- **Pros**: Easy to understand; Resembles the *p*-value - $r \simeq -1.$ *
- **Cons**: like the *p*-values, a *low* p_d cannot be used to support the null.

p-MAP

- The *density ratio* between the null and the MAP value.
- Values range from 1 (the null *is* the MAP) to ~0 (the MAP is much much more probable than the null).

```
p_map(diffs_Congruency)
```

```
## # MAP-based p-value
##
## Parameter    | p_MAP
## -------------------
## Interference | 0.086
## Facilitation | 0.651
```



For the Interference effect it seems like the MAP is more th 10 times more probable than the null. But for the Facilitation effect it is not even twice as probable.

- **Pros**: Closely related to LRT tests - familiar; Also closely associated with the *p*-value.
- **Cons**: Again, a *high p*-MAP cannot be used to support the null.

p-ROPE

- The probability that our estimate is *basically* null.
- We first define a **Region of Practical Equivalence (ROPE)** - a range of effects that are, for any practical purposes, the same as no effect at all.

For the Congruency effects, we will define any effect that is smaller in magnitude than 30ms, to be consider to be just as good as no effect at all - so ROPE [-30, +30].

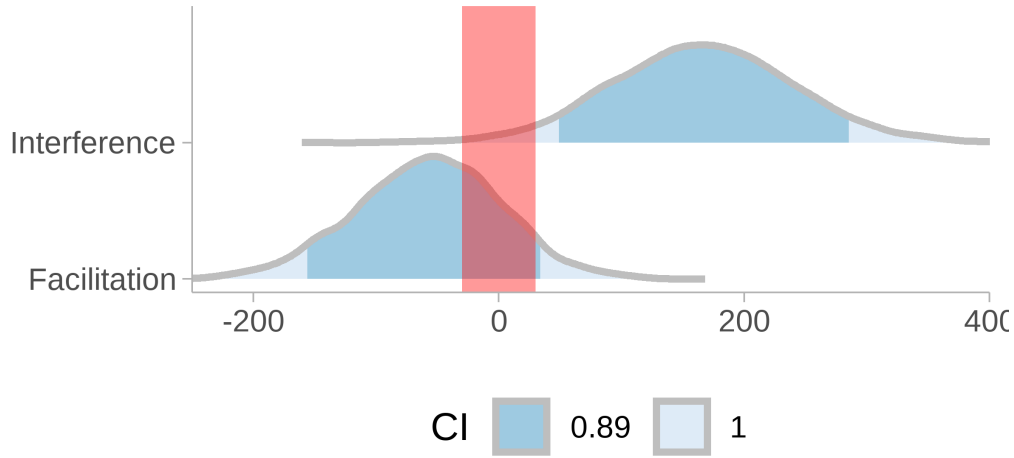
We can also have a one sided ROPE, with [-Inf, +30], etc.

p-ROPE

- How much of the posterior falls in the ROPE.
  - Or: How much of the most probable values (e.g., those in the HDI) fall in the ROPE.

```
rope(diffs_Congruency,
     range = c(-30, 30), ci = 0.89)
```

```
## # Proportion of samples inside the ROPE [-30.00, 3
##
## Parameter    | inside ROPE
## --------------------------
## Interference |      0.00 %
## Facilitation |     29.99 %
```



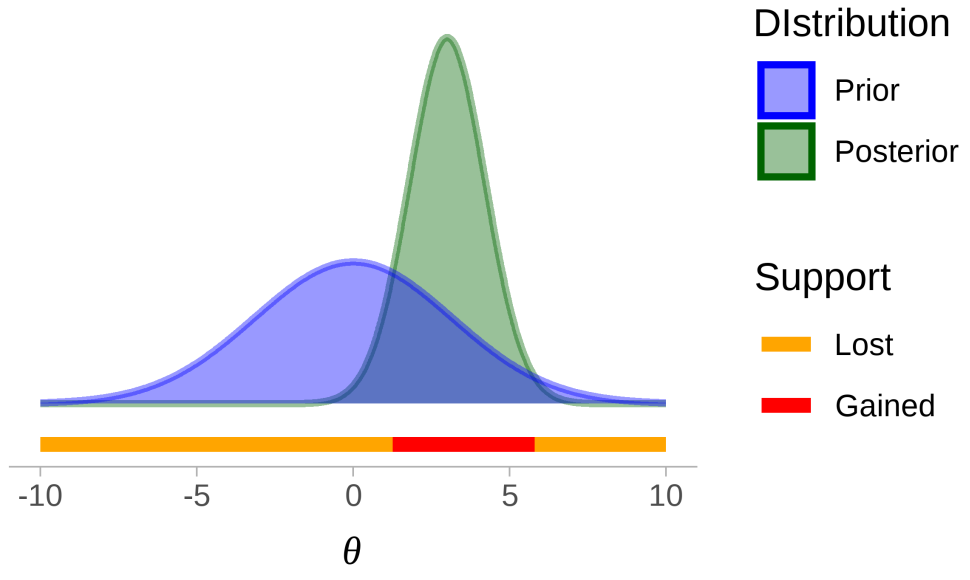
It is very improbable that the Interference effect is very small. *But* there is about a 30% that among 4 year olds, there is no Facilitation effect - (though not very conclusive) we are supporting the null!

The p_d, *p*-MAP and ROPE are **posterior based methods** - they inform us about the accumulated information in the priors + our data.

Often we are interested in **what has been *learned* in the current study, from the current data**.

E.g., by now it's clear that there exists an Interference effect in Flanker's task. But **which values of the effect are supported or contradicted by the *current* data?** Maybe our data supports the null value(s) - what can be learnt from that?

To answer these types of questions we can *compare* The Prior to The Posterior to see what our data taught us - what values became more / less plausible.



We can use this information to look at different *sets of parameter values* - or **hypotheses** - E.g. $H_{small}: \theta \in [-3, 3]$, $H_{positive}: \theta \in [0, \infty]$ and ask:

> Which *hypothesis* is supported *more* by the data?

The Bayes Factor

This index of evidence is a ***Bayes Factor***:

- It quantifies how the prior was *updated* to the posterior.
- It compares two "hypotheses".

Any measure that quantifies this 👆 is a Bayes factor.

There are many different type of questions that can be answered with Bayes factors - we will be looking at two.

For technical reasons we need a model that represents *only our priors* - which we will then *compare* to the results from our updated (posterior) model.

We can do that with the `unupdate()` function:

```
# Get the priors only ("un-update" the model).
m_flanker_prior <- unupdate(m_flanker)
```

The Null-Interval Bayes Factor

The null-interval Bayes factor is an extension of the ROPE test;

> How has the *relative probability*[1] of the the effect being practically null changed? Does the data support or contradict the effect being null?

The two hypotheses we will be comparing, using the same ROPE:

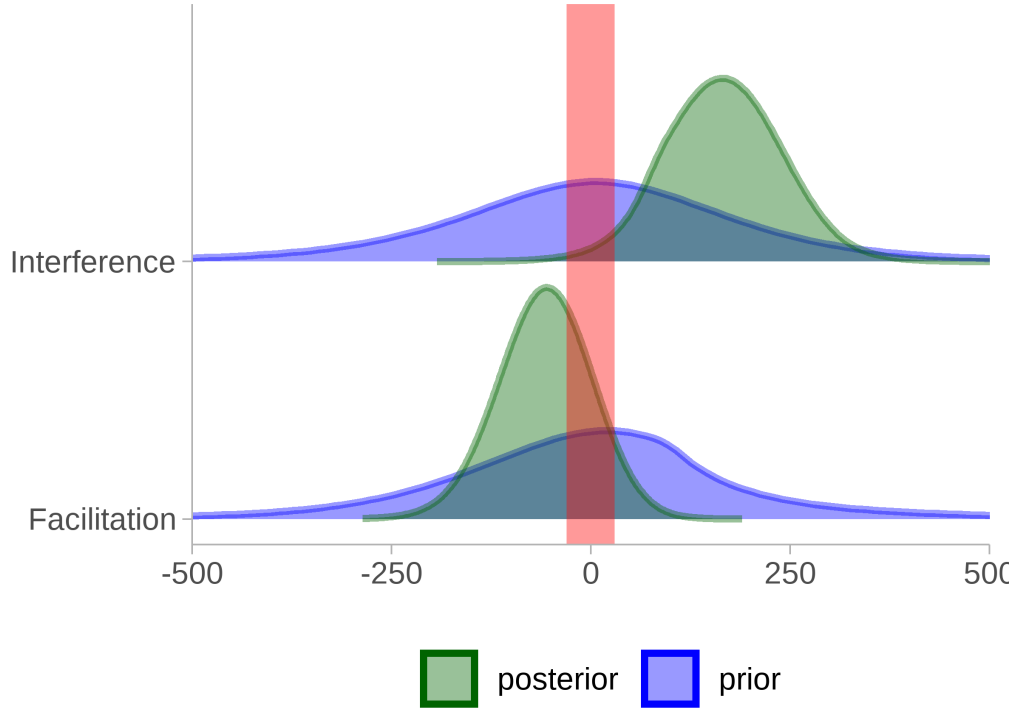
- $H_0 : \text{effect} \in [-30, +30]$
- $H_A : \text{effect} \notin [-30, +30]$
  - Or: $H_A : \text{effect} < -30 \text{ or } +30 < \text{effect}$

[1] The odds of the effect being inside the ROPE to it being outside the ROPE.

The Null-Interval Bayes Factor

```
bayesfactor_parameters(
  diffs_Congruency,
  prior = m_flanker_prior,
  null = c(-30, 30) # same ROPE as before
)
```

```
## # Bayes Factor (Null-Interval)
##
## Parameter    |    BF
## --------------------
## Interference | 6.022
## Facilitation | 0.518
##
## * Evidence Against The Null: [-30, 30]
```



- For the Interference effect, the ROPE has *become* relatively less probable - with the data giving 6 times more support for non-ROPE values.
- For the Facilitation effect, the ROPE has *become* relatively **more** probable - with the data giving (1/0.5 =) 2 times more support compared to the non-ROPE values.

The Point-Null Bayes Factor

The point-null can be thought of as the null-interval Bayes factor with an infinitesimally small ROPE - that includes only one null value, exactly.

> How has the probability[1,2] of the the null value changed? Does the data support or contradict the effect being null?

This Bayes factor is also called the *Savage-Dickey density ratio*, and it is analogous to a Bayes factor comparing two nested models.

The two hypotheses we will be comparing:

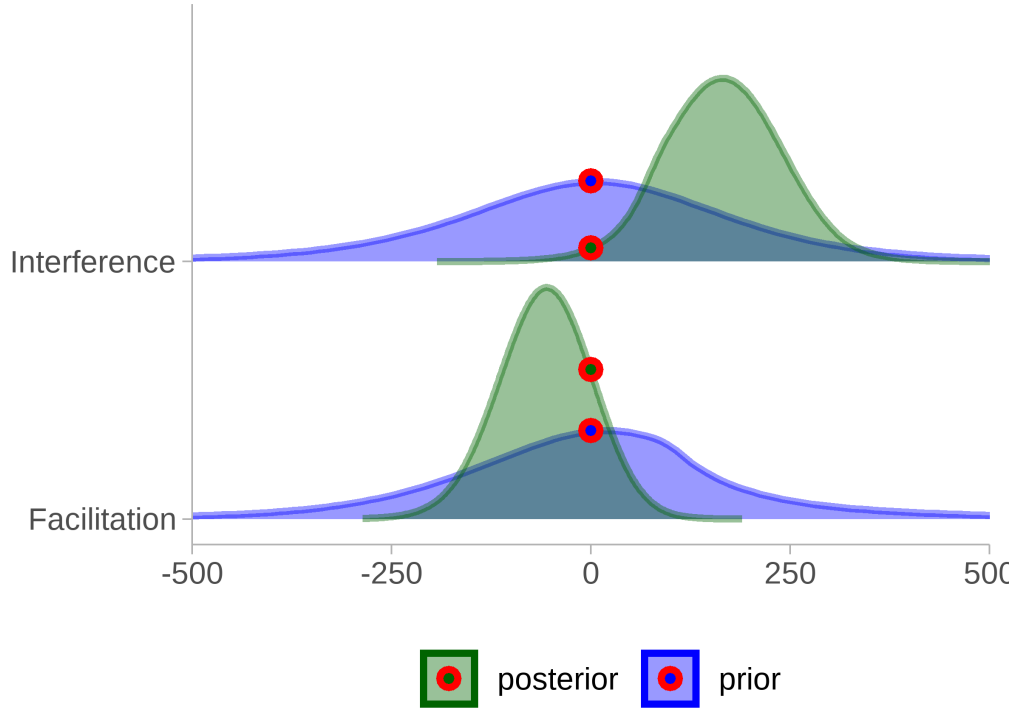
- $H_0 : \text{effect} = 0$
- $H_A : \text{effect} \neq 0$
  - Or: $H_A : \text{effect} < 0 \text{ or } 0 < \text{effect}$

[1] Actually the density of the null.
[2] This is also relative - if the null became more probable, necessarily the non-null values became less, and vice versa.

The Point-Null Bayes Factor

```
bayesfactor_parameters(
  diffs_Congruency,
  prior = m_flanker_prior,
  null = 0
)
```

```
## # Bayes Factor (Savage-Dickey density ratio)
##
## Parameter    |    BF
## --------------------
## Interference | 5.930
## Facilitation | 0.591
##
## * Evidence Against The Null: [0]
```



- For the Interference effect, the *mass* of the posterior is shifted *away* from the null (compared to the prior) - the data giving ~6 times more support for non-null values.
- For the Facilitation effect the mass has moved *towards* 0, the data giving (1/0.6 =) 1.7 times more support compared to the non-null values.

The Point-Null Bayes Factor

```
bayesfactor_parameters(
  diffs_Congruency,
  prior = m_flanker_prior,
  null = 0
)
```

```
## # Bayes Factor (Savage-Dickey density ratio)
##
## Parameter    |    BF
## --------------------
## Interference | 5.930
## Facilitation | 0.591
##
## * Evidence Against The Null: [0]
```

```
bayesfactor_parameters(
  diffs_Congruency,
  prior = m_flanker_prior,
  null = c(-30, 30) # same ROPE as before
)
```

```
## # Bayes Factor (Null-Interval)
##
## Parameter    |    BF
## --------------------
## Interference | 6.022
## Facilitation | 0.518
##
## * Evidence Against The Null: [-30, 30]
```

Here the point-null and the null-interval BFs gave similar results, but that need not be the case - depending on the effect size, the definition of the ROPE, the sample size, etc.

Other Bayes Factors

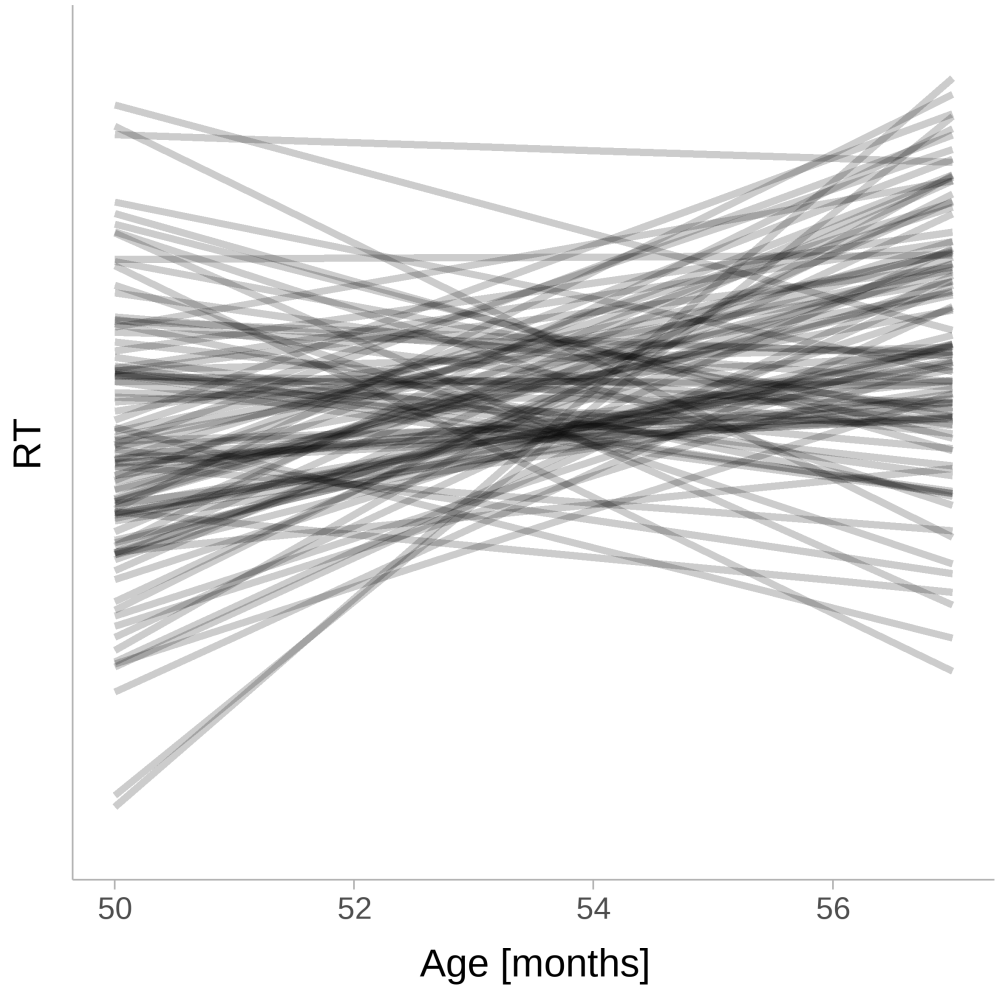
- **Directional** null-interval / point-null Bayes factors
  - e.g., [-30, +30] *vs* [+30, Inf]
- Bayes factor for **dividing hypotheses**
  - e.g., [-Inf, 0] *vs* [0, Inf]
- **Model restricted** Bayes factors
  - [Incongruent > Neutral > Congruent] vs [Incongruent ≠ Neutral ≠ Congruent]
- And more...

Read more about these Bayes factors here!

Age

For *covariates*, we can present the posterior distribution of slopes, but be can also present a *trace plot* of slopes from the posterior.

For example, we can sample 100 slopes from the posterior, and plot each one:



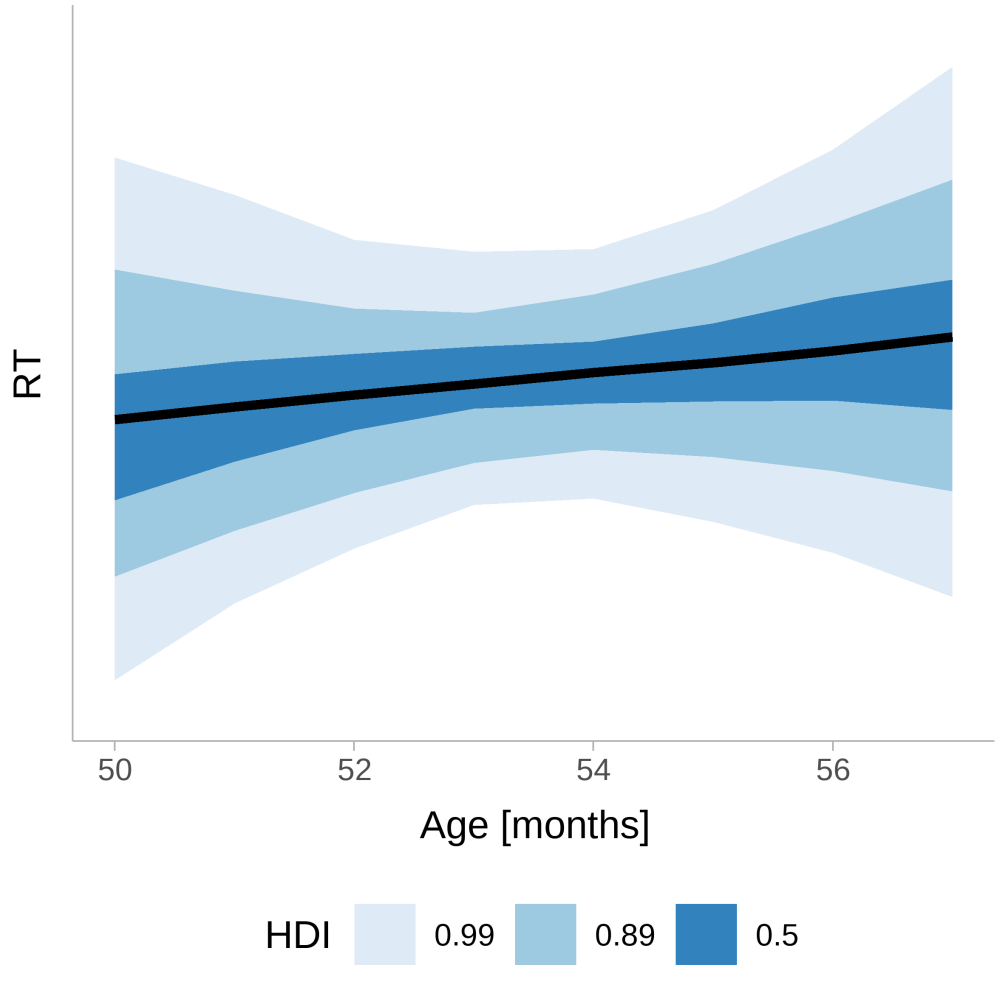
Age

Here too we can summarize the posterior distribution:

```
slope_age <- emtrends(m_flanker, ~1, "age_mo")

describe_posterior(slope_age, test = NULL)
```

```
## # Description of Posterior Distributions
##
## Parameter | Median |              89% CI
## ------------------------------------------
## overall   | 22.476 | [-47.475, 101.115]
```



p-Direction & *p*-MAP

```
p_direction(slope_age)
```

```
## # Probability of Direction (pd)
##
## Parameter |     pd
## ------------------
## overall   | 67.70%
```

```
p_map(slope_age)
```

```
## # MAP-based p-value
##
## Parameter | p_MAP
## -----------------
## overall   | 0.830
```

Not very decisive... (remember, these cannot be used to support the null!)

p-ROPE

For the ROPE - I think any effect smaller an overall change of less than 500ms a year = ~40ms a month, is practically 0 (you may disagree...):

```
rope(slope_age, range = c(-40, 40), ci = 0.89)
```

```
## # Proportion of samples inside the ROPE [-40.00, 40.00]:
##
## Parameter | inside ROPE
## ----------------------
## overall   |     63.89 %
```

There is about a 60% probability that the effect of age on reaction times is *practically* nothing!

Not strongly conclusive, but at the very least it is suggestive!

Bayes Factor

```
bayesfactor_parameters(
  slope_age,
  prior = m_flanker_prior,
  null = 0
)
```

```
## # Bayes Factor (Savage-Dickey density ratio)
##
## Parameter |    BF
## -----------------
## overall   | 0.052
##
## * Evidence Against The Null: [0]
```

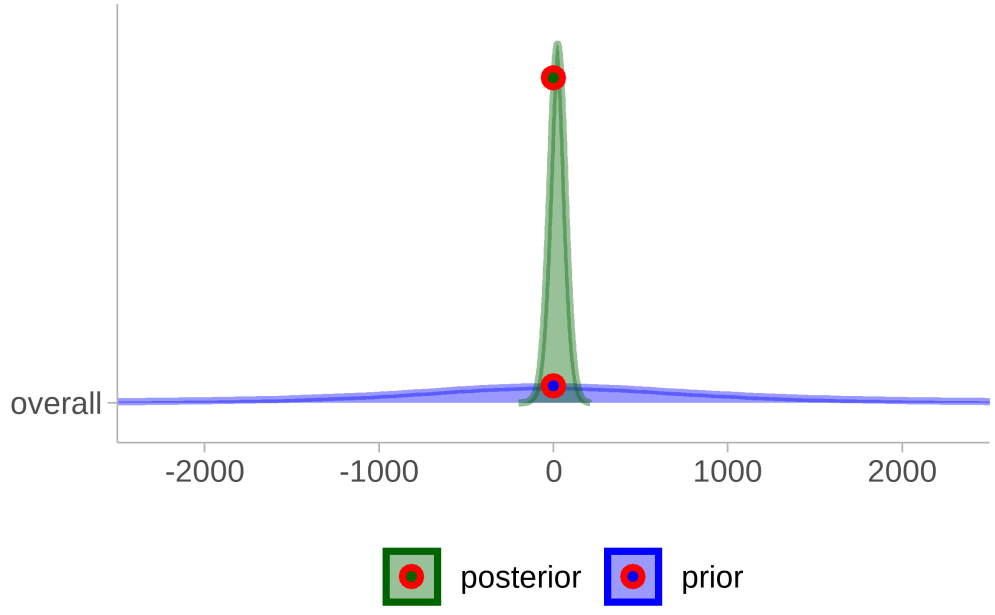
Wow! It seems that the data strongly support (by a factor of 1/0.05 = 20) the effect of age being null over it being non-null!

But wait - the Bayes factor measures the change from the prior to the posterior... But what was our prior here?

Bayes Factor

```
bayesfactor_parameters(
  slope_age,
  prior = m_flanker_prior,
  null = 0
)
```

```
## # Bayes Factor (Savage-Dickey density ratio)
##
## Parameter |    BF
## -----------------
## overall   | 0.052
##
## * Evidence Against The Null: [0]
```



We used a super vague prior - which give some non-trivial probability to extreme effects!

So is it really surprising that the posterior is now, relatively closer to the *null*? ***No.***

With wide and uninformative enough priors, the Bayes factor will **always favor the null / ROPE**!
DO NOT COMPUTE BAYES FACTORS WITH UNINFORMATIVE PRIORS! *

Recommendations

What to actually report?

We (Makowski et al., 2019) recommend reporting for inferential statistics:

- **The *p*-direction**: Easy to understand, easy to "translate" to *p*-values.
- ***p*-ROPE**: Provides information about the practical relevance of the effect, and allows to accept the null.

If informed priors are used,

- **Bayes factor** (instead or in addition to the *p*-ROPE): Provides information about hypotheses supported or contradicted by the data.

Summary

What you now know! 🎓

- What a Bayesian model *is*.
- What Bayes can give you, that no one else can.
- A taste of Bayesian model fitting with `brms`.
- The richness of inferences that can be made with Bayesian statistics.

Suggested Reading

For Bayesian Beginners

- Makowski, D., Ben-Shachar, M. S., Chen, S. H., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in psychology, 10*, 2767.
    - `bayestestR` guides and articles.
- Van de Schoot, R. et al (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers, 1*(1), 1-26.
- Bayesian Inference for Psychology. *Psychonomic Bulletin and Review*.

Books

- Kruschke, J. (2014). Doing bayesian data analysis: A tutorial with r, jags, and stan. Academic Press.
- McElreath, R. (2018). Statistical rethinking: A bayesian course with examples in r and stan. Chapman; Hall/CRC.
    - Richard's YouTube channel

Thank you!

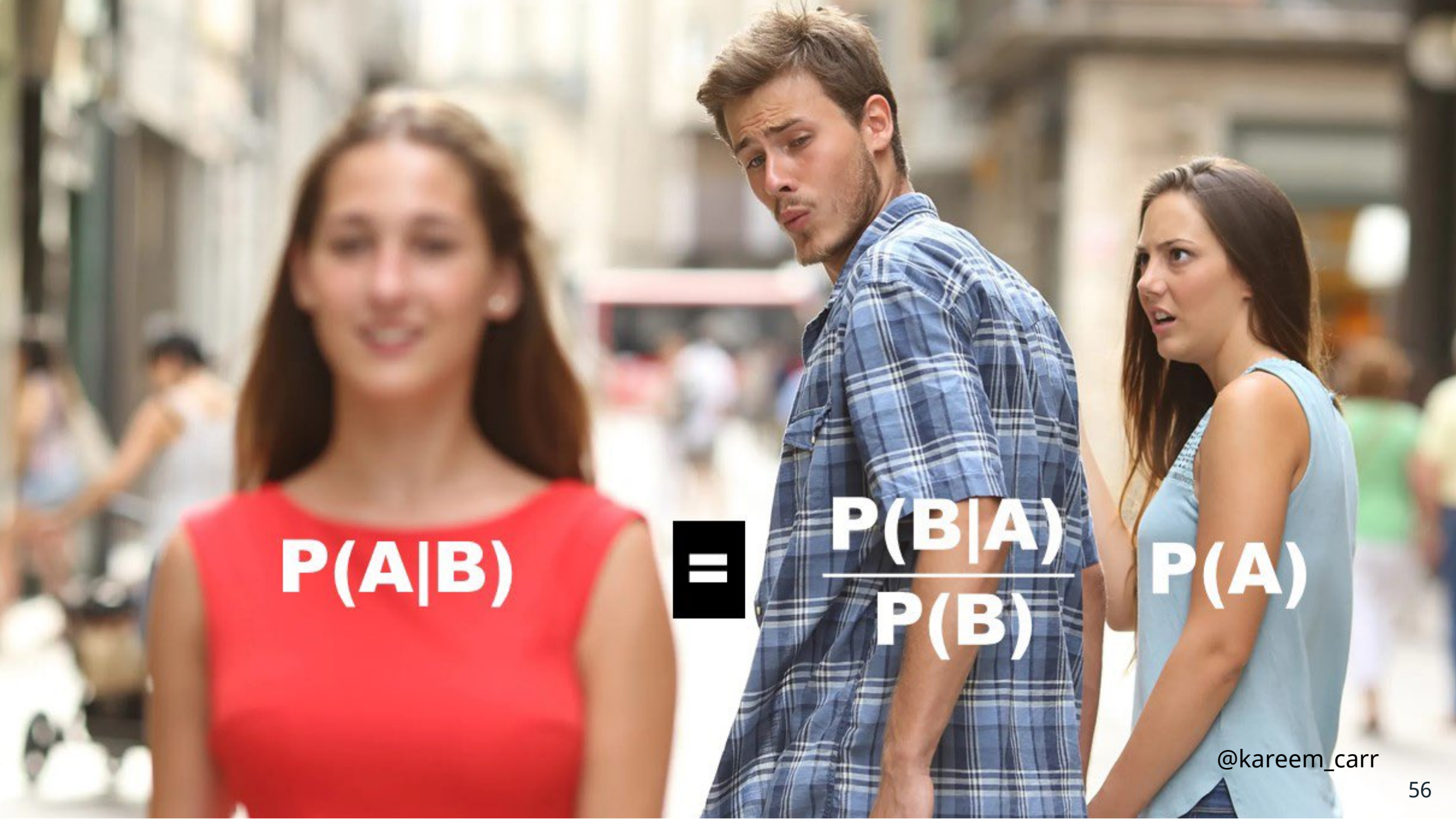
Follow me!

@mattansb | @mattansb | Blog

Slides created with the R package ***xaringan***.

The **bayestestR** package is part of the easystats project.
Core team members:

- Me 👋
- Dominique Makowski (@Dom_Makowski)
- Daniel Lüdecke (@strengejacke)
- Indrajeet Patil (@patilindrajeets)


P(A|B)
=
P(B|A)
P(B)
P(A)
@kareem_carr