# Bayes factors

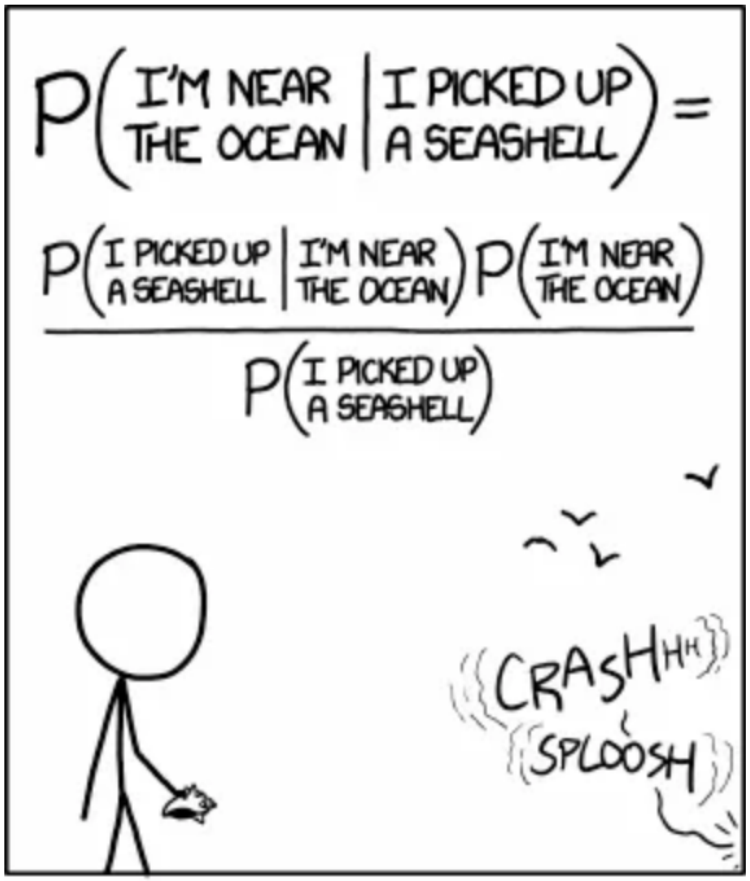
Nicole Cruz

SMLP 2025

Last update: August 26, 2025

# Contents

# From Bayes theorem to Bayes factors




https://xkcd.com/1236/

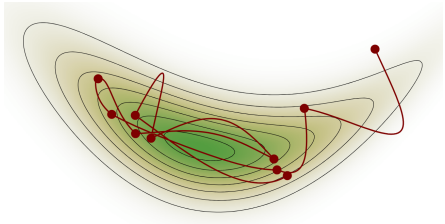
# Bayes theorem

Follows directly from the axioms of (classical) probability theory

- $P(D \,\&\, H) = P(D)P(H|D)$
- $P(D \,\&\, H) = P(H)P(D|H)$
- $P(D)P(H|D) = P(H)P(D|H)$
- $P(H|D) = P(H)P(D|H) \,/\, P(D)$.

(Joyce, 2003).

# Bayesian conditionalization

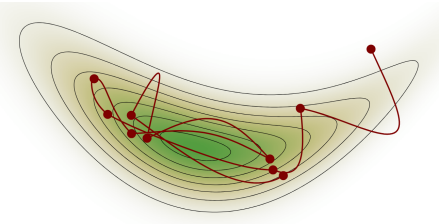


Generalizes Bayes theorem from static to dynamic setting, allowing inference from prior to posterior state of affairs.

- Bayes theorem: $P_1(H|D) = \frac{P_1(H)P_1(D|H)}{P_1(D)}$.
- Bayesian conditionalization: $P_2(H) = P_1(H|D)$.
  - $P_1$ = probability at time 1; $P_2$ = probability at time 2.
  - Presupposes the data D is learned with certainty (more general Jeffrey conditionalization does not make this assumption).
  - Presupposes invariance in the conditional probabilities between time points: $P_1(H|D) = P_2(H|D)$.
  - When assumptions met and priors internally consistent, then repeated application of Bayesian conditionalization tends to lead to convergence with the ground truth (where there is one).

(Hájek, 2023; Joyce, 2003).

# Bayesian conditionalization



Generalizes Bayes theorem from static to dynamic setting, allowing inference from prior to posterior state of affairs.

- Bayes theorem: $P_1(H|D) = \frac{P_1(H)P_1(D|H)}{P_1(D)}$.
- Bayesian conditionalization: $P_2(H) = P_1(H|D)$.
  - $P_1$ = probability at time 1; $P_2$ = probability at time 2.
  - Presupposes the data D is learned with certainty (more general Jeffrey conditionalization does not make this assumption).
  - Presupposes invariance in the conditional probabilities between time points: $P_1(H|D) = P_2(H|D)$.
  - When assumptions met and priors internally consistent, then repeated application of Bayesian conditionalization tends to lead to convergence with the ground truth (where there is one).

(Hájek, 2023; Joyce, 2003).

# Data can be added sequentially


- **Posterior**
- Likelihood

$$P_2(H) = P_1(H|D) = \frac{P_1(D|H)\,P_1(H)}{P_1(D)}$$

- Prior
- Marginal likelihood

# Data can be added sequentially


- Posterior
- **Likelihood**

$$P_2(H) = P_1(H|D) = \frac{P_1(D|H)\ P_1(H)}{P_1(D)}$$

- Prior
- Marginal likelihood

# Data can be added sequentially


- Posterior
- Likelihood

$$P_2(H) = P_1(H|D) = \frac{P_1(D|H)\ P_1(H)}{P_1(D)}$$

- **Prior**
- Marginal likelihood

# Data can be added sequentially


- Posterior
- Likelihood

$$P_2(H) = P_1(H|D) = \frac{P_1(D|H)\ P_1(H)}{P_1(D)}$$

- Prior
- **Marginal likelihood**

# Data can be added sequentially


- Posterior
- Likelihood

$$P_3(H) = P_1(H|D_2, D_1) = \frac{P_1(D_2|H)P_1(D_1|H)\ P_1(H)}{P_1(D_2)P_1(D_1)}$$

- Prior
- Marginal likelihood

# Discrete vs. continuous

Same idea, slightly different notation

$$\text{Discrete: } P(H|D) = \frac{P(D|H)P(H)}{P(D) = \sum_H P(D|H)P(H)}$$

$$\text{Continuous: } P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y) = \int_\theta P(Y|\theta)P(\theta)d\theta}$$

# One or more parameters

- In models with several parameters, the posterior distribution is multidimensional and $\theta$ is a vector of parameters:

$$\underbrace{\mathbf{P}(\boldsymbol{\theta}|\mathbf{Y})}_{\textit{Puposterior}} = \underbrace{\mathbf{P}(\boldsymbol{\theta})}_{\text{Prior}} \times \underbrace{\frac{\mathbf{P}(\mathbf{Y}|\boldsymbol{\theta})}{\mathbf{P}(\mathbf{Y})}}_{\text{Updating factor}}$$

(Veenman et al., 2023).

# Joint vs. marginal

- Often we are not interested in the joint posterior vector $\theta$ with all parameters, but only in the posterior of a single parameter, e.g. $\mu$. In this case we seek the marginal posterior for $\mu$.
- We obtain the marginal posterior for a parameter by integrating out the remaining parameters from the joint posterior. For $\mu_\beta$ (where $\theta$ now represents all parameter except $\mu_\beta$):

(Veenman et al., 2023).

# Joint vs. marginal

- Often we are not interested in the joint posterior vector $\theta$ with all parameters, but only in the posterior of a single parameter, e.g. $\mu$. In this case we seek the marginal posterior for $\mu$.
- We obtain the marginal posterior for a parameter by integrating out the remaining parameters from the joint posterior. For $\mu_\beta$ (where $\theta$ now represents all parameter except $\mu_\beta$):

$$\mathbf{P}(\mu_\beta|\mathbf{Y}) = \int \mathbf{P}(\mathbf{Y}|\mu_\beta, \boldsymbol{\theta})\mathbf{P}(\mu_\beta)\mathbf{P}(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}$$

(Veenman et al., 2023).

# Marginal for a model $M$

$$\text{Posterior} = P(\theta|Y,M) = \frac{P(Y|\theta,M)P(\theta|M)}{P(Y|M)} = \frac{\text{likelihood x prior}}{\text{marginal likelihood}}$$

$$P(Y|M)_{discrete} = \sum_{i=1}^{k} P(Y|\theta_i,M)P(\theta_i|M)$$

$$P(Y|M)_{continuous} = \int_{\Theta} P(Y|\theta,M)P(\theta|M)d\theta$$

- The marginal likelihood is the distribution of the *data* $Y$ *given model* $M$, averaged across all parameters $\theta$.

(Lee & Wagenmakers, 2013, Nicenboim, Schad, & Vasishth, 2025; Veenman et al., 2023).

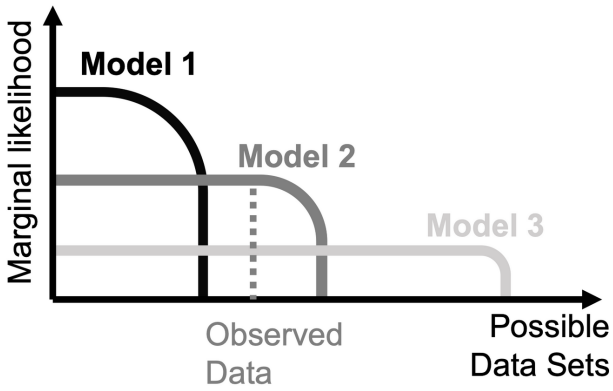
# Some uses

$$P(Y|M) = \int_{\Theta} P(Y|\theta, M)P(\theta|M)d\theta$$

- Prior to observing the data, the marginal likelihood can be used to predict their distribution.
- After data are observed, the marginal likelihood can be used to assess the correspondence between data and model predictions.
- It is not always easy to interpret on its own - but helpful to compare it to the marginal likelihoods of alternative models for the same data.

(Nicenboim, Schad, & Vasishth, 2025; Veenman et al., 2023).

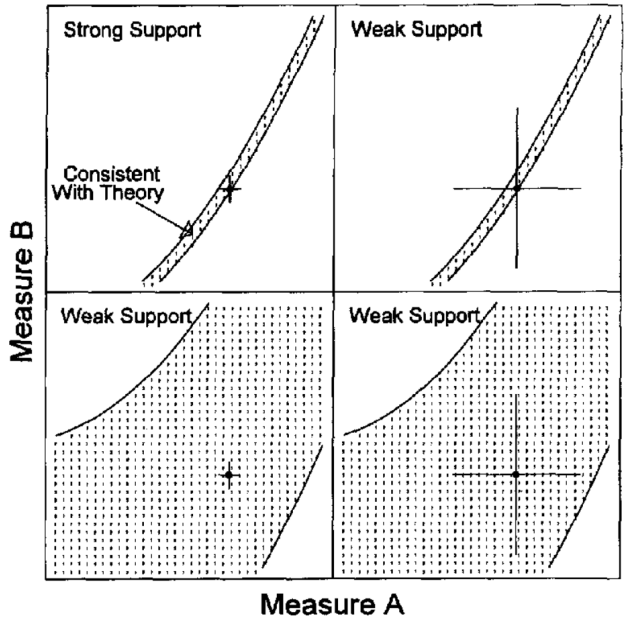
# Model flexibility




- A model has high marginal likelihood if it makes high proportion of good predictions.
- Model predictions are normalized: the total probability that models assign to different data patterns is the same for all models.
- A model that predicts many different outcomes spreads its prior predictive density more thinly among them. This flexibility gives the model higher chances of predicting the actually observed outcome; but due to normalization it cannot predict it with high probability.

(Lee & Wagenmakers, 2013, Nicenboim, Schad, & Vasishth, 2025; Veenman et al., 2023).

# Complexity and informativeness




- More complex, flexible models are less informative and more difficult to falsify $\rightarrow$ less useful as scientific theories.
- Models are more complex when they have (a) wider priors, (b) more parameters, (c) less constrained functions.
- Bayesian model comparison inherently penalizes unnecessarily complex models (Occam's razor).

(Roberts & Pashler, 2000).

# Ratios of likelihoods: Bayes factors

$$BF_{12} = \frac{P(Y|M_1)}{P(Y|M_2)}$$

- $BF_{12}$ is the ratio of the marginal likelihood of model $M_1$ to the marginal likelihood of Model $M_2$.
- Corresponds to the relative evidence the data provide for $M_1$ relative to $M_2$.
- Values higher than 1 provide evidence for $M_1$, values below 1 provide evidence for $M_2$, and values close to 1 provide inconclusive evidence.
- Does not depend on a specific parameter value - all possible prior parameter values are taken into account simultaneously.

(Nicenboim, Schad, & Vasishth, 2025;Veenman et al., 2023).

# Ratios of likelihoods: Bayes factors

$$BF_{12} = \frac{P(Y|M_1)}{P(Y|M_2)}$$

- $BF_{12}$ is the ratio of the marginal likelihood of model $M_1$ to the marginal likelihood of Model $M_2$.
- Corresponds to the relative evidence the data provide for $M_1$ relative to $M_2$.
- Values higher than 1 provide evidence for $M_1$, values below 1 provide evidence for $M_2$, and values close to 1 provide inconclusive evidence.
- Does not depend on a specific parameter value - all possible prior parameter values are taken into account simultaneously.

(Nicenboim, Schad, & Vasishth, 2025;Veenman et al., 2023).

# Ratios of likelihoods: Bayes factors

$$BF_{12} = \frac{P(Y|M_1)}{P(Y|M_2)}$$

- $BF_{12}$ is the ratio of the marginal likelihood of model $M_1$ to the marginal likelihood of Model $M_2$.
- Corresponds to the relative evidence the data provide for $M_1$ relative to $M_2$.
- Values higher than 1 provide evidence for $M_1$, values below 1 provide evidence for $M_2$, and values close to 1 provide inconclusive evidence.
- Does not depend on a specific parameter value - all possible prior parameter values are taken into account simultaneously.

(Nicenboim, Schad, & Vasishth, 2025;Veenman et al., 2023).

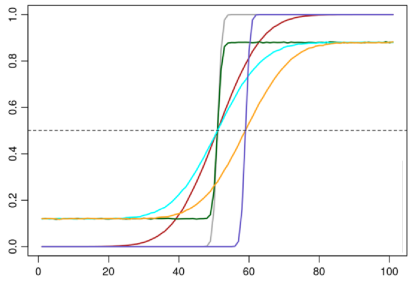
# Ratios of likelihoods: Bayes factors

$$BF_{12} = \frac{P(Y|M_1)}{P(Y|M_2)}$$

- $BF_{12}$ is the ratio of the marginal likelihood of model $M_1$ to the marginal likelihood of Model $M_2$.
- Corresponds to the relative evidence the data provide for $M_1$ relative to $M_2$.
- Values higher than 1 provide evidence for $M_1$, values below 1 provide evidence for $M_2$, and values close to 1 provide inconclusive evidence.
- Does not depend on a specific parameter value - all possible prior parameter values are taken into account simultaneously.

(Nicenboim, Schad, & Vasishth, 2025;Veenman et al., 2023).

# Strength of evidence




- Cutoff values have been proposed to categorise *BF*s according to the strength of evidence they provide.
- These are just suggestions for when no more specific evidence strength criteria (informed by research questions and goals) are available.
- Some Bayesians reject cutoff values altogether, preferring to report the quantitative evidence provided by the *BF* directly.
- Alternative: combine the continuous evidence of *BF*s with discrete decision thresholds to maximise expected utility of acting on a particular hypothesis.

(Nicenboim, Schad, & Vasishth, 2025;Veenman et al., 2023).

# Jeffrey's evidence thresholds

**Table 7.1** Evidence categories for the Bayes factor $BF_{12}$ (Jeffreys, 1961).

| Bayes factor $BF_{12}$ | | | Interpretation |
|---|---|---|---|
| | $>$ | 100 | Extreme evidence for $\mathcal{M}_1$ |
| 30 | – | 100 | Very strong evidence for $\mathcal{M}_1$ |
| 10 | – | 30 | Strong evidence for $\mathcal{M}_1$ |
| 3 | – | 10 | Moderate evidence for $\mathcal{M}_1$ |
| 1 | – | 3 | Anecdotal evidence for $\mathcal{M}_1$ |
| | 1 | | No evidence |
| 1/3 | – | 1 | Anecdotal evidence for $\mathcal{M}_2$ |
| 1/10 | – | 1/3 | Moderate evidence for $\mathcal{M}_2$ |
| 1/30 | – | 1/10 | Strong evidence for $\mathcal{M}_2$ |
| 1/100 | – | 1/30 | Very strong evidence for $\mathcal{M}_2$ |
| | $<$ | 1/100 | Extreme evidence for $\mathcal{M}_2$ |

(Lee & Wagenmakers, 2013).

# Some properties of Bayes factors

$$BF_{1,2} = \frac{P(Y|M_1)}{P(Y|M_2)}$$

$$BF_{2,1} = 1/BF_{1,2}$$

$$BF_{a,c} = BF_{a,b} * BF_{b,c}$$

- *BFs* offer quantitative evidence for one model relative to another.
- The evidence is not symmetrical ($BF_{a,b} \neq BF_{b,a}$); but knowing the evidence for $a$ relative to $b$ is enough to compute the evidence for $b$ relative to $a$.
- *BFs* are transitive.
- *BFs* are odds ratios (OR), not probabilities.

(Rouder & Morey, 2018; Veenman et al., 2023).

# From *BFs* back to Bayes theorem

$$\frac{P(M1|y)}{P(M2|y)} = \frac{P(M1)}{P(M2)} x \frac{P(y|M1)}{P(y|M2)}$$

$$\text{Posterior odds}_{12} = \text{Prior odds}_{12} \, x \, BF_{12}$$

- *BF* indicates amount by which we should update our relative belief between the two models in light of data and priors.
- To find out which model has the highest posterior probability, we need to combine the *BF* with the relative prior probability of each model.

(Nicenboim, Schad, & Vasishth, 2025).

# Methods for computing Bayes factors

# Two main estimation algorithms

- Savage-Dickey density ratio
- Bridge sampling

(Veenman et al., 2023).

# Savage-Dickey density ratio (SD density-ratio)

- Estimates the mean effect of a parameter by comparing nested models with vs. without the parameter.
  - $\rightarrow$ E.g. $M_1$ estimates the distribution of a parameter, while in $M_2$ the parameter is set to zero or to another constant.
- Die models compared include the same effects of interindividual differences, making them similar to type III sum os squares in ANOVA.
- Relatively efficient, computable analytically in special cases.
- In `brms`: e.g. via `hypothesis()` function of the same package.

(Veenman et al., 2023; Wagenmakers et al., 2010).

# Savage-Dickey density ratio (SD density-ratio)

- Estimates the mean effect of a parameter by comparing nested models with vs. without the parameter.
  - $\rightarrow$ E.g. $M_1$ estimates the distribution of a parameter, while in $M_2$ the parameter is set to zero or to another constant.
- Die models compared include the same effects of interindividual differences, making them similar to type III sum os squares in ANOVA.
- Relatively efficient, computable analytically in special cases.
- In `brms`: e.g. via `hypothesis()` function of the same package.

(Veenman et al., 2023; Wagenmakers et al., 2010).

# SD density-ratio: limitations

- Estimation quality depends on the density of the parameter distribution at the point of comparison (e.g. zero). If the distribution has low density at that point (e.g. when the effect of the parameter is large) then its estimation becomes unreliable.
  - $\rightarrow$ Can be partially compensated for by drawing more samples
- Use limited to nested models that differ in a single parameter.

(Veenman et al., 2023).

# SD density-ratio: limitations

- Estimation quality depends on the density of the parameter distribution at the point of comparison (e.g. zero). If the distribution has low density at that point (e.g. when the effect of the parameter is large) then its estimation becomes unreliable.
  $\rightarrow$ Can be partially compensated for by drawing more samples
- Use limited to nested models that differ in a single parameter.

(Veenman et al., 2023).

# Bridge sampling

- More general and versatile than SD density-ratio.
- Applicable to non-nested models and to models that differ in more than one parameter.
- Based on sampling the posterior distribution of a model to obtain the corresponding marginal likelihood of the data.
- The marginal likelihoods of the data given different models can then be directly compared.
- In brms: e.g. via `bayes_factor()` from `multibridge` package, or `bayesfactor_models()` from `bayestestR` package.

(Gronau et al., 2017; Veenman et al., 2023).

# Bridge sampling

- More general and versatile than SD density-ratio.
- Applicable to non-nested models and to models that differ in more than one parameter.
- Based on sampling the posterior distribution of a model to obtain the corresponding marginal likelihood of the data.
- The marginal likelihoods of the data given different models can then be directly compared.
- In `brms`: e.g. via `bayes_factor()` from `multibridge` package, or `bayesfactor_models()` from `bayestestR` package.

(Gronau et al., 2017; Veenman et al., 2023).

# Bridge sampling: Limitations

- Less efficient than SD density ratio.
- Like all sampling based methods it offers an estimate, not an exact value. *BF* estimation typically requires more samples than the estimation of posterior model parameters.
  - → Rule of thumb: 10 x as many iterations
  - → Stability analysis by repeatedly running the model, and for each model run repeatedly computing the *BF* (e.g. 5 x 5 times).

(Veenman et al., 2023).

# Bridge sampling: Limitations

- Less efficient than SD density ratio.
- Like all sampling based methods it offers an estimate, not an exact value. *BF* estimation typically requires more samples than the estimation of posterior model parameters.
  - → Rule of thumb: 10 x as many iterations
  - → Stability analysis by repeatedly running the model, and for each model run repeatedly computing the *BF* (e.g. 5 x 5 times).

(Veenman et al., 2023).

# Bridge sampling: Limitations

- Less efficient than SD density ratio.
- Like all sampling based methods it offers an estimate, not an exact value. *BF* estimation typically requires more samples than the estimation of posterior model parameters.
  - $\rightarrow$ Rule of thumb: 10 x as many iterations
  - $\rightarrow$ Stability analysis by repeatedly running the model, and for each model run repeatedly computing the *BF* (e.g. 5 x 5 times).

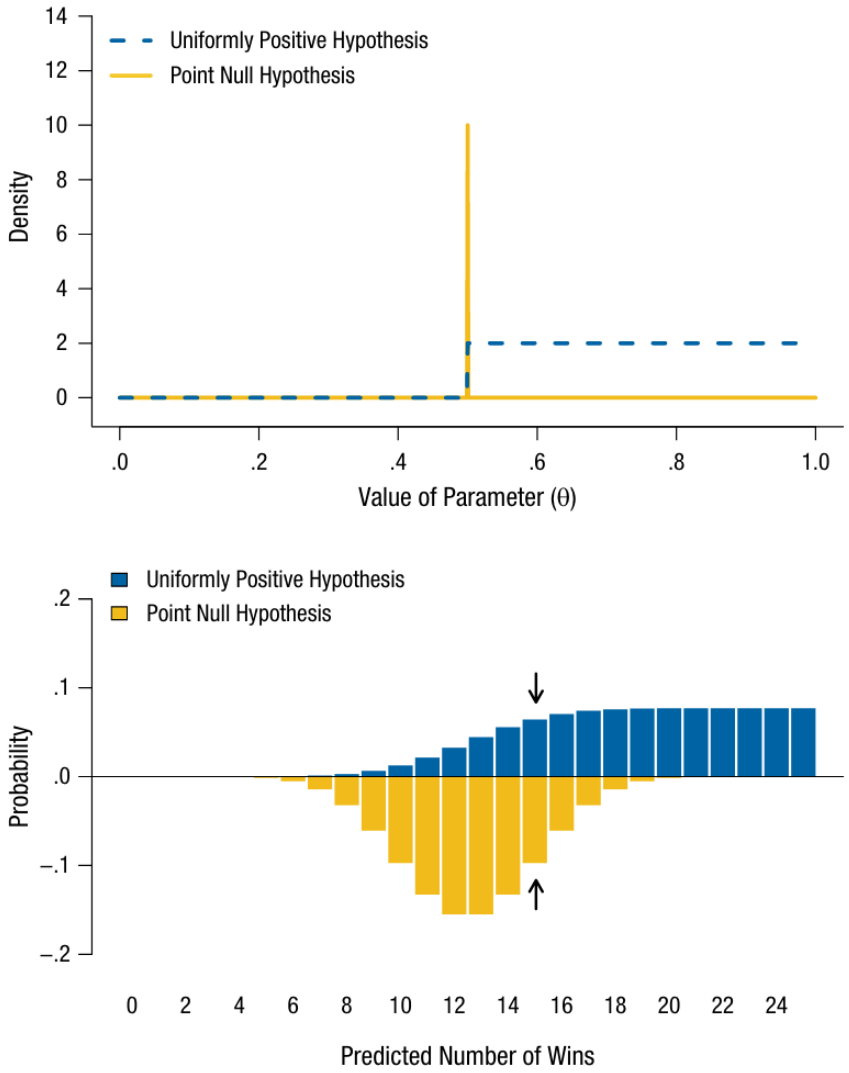
(Veenman et al., 2023).

# Types of model comparison

# Wide range of possible comparisons

- Nested or non-nested models when applied to the same data
- Evidence for Null vs. for alternative Hypotheses
- Point vs. range of values
- Truncated vs. non-truncated distributions
- Normal or other distribution forms
- Parameters for means or for variances

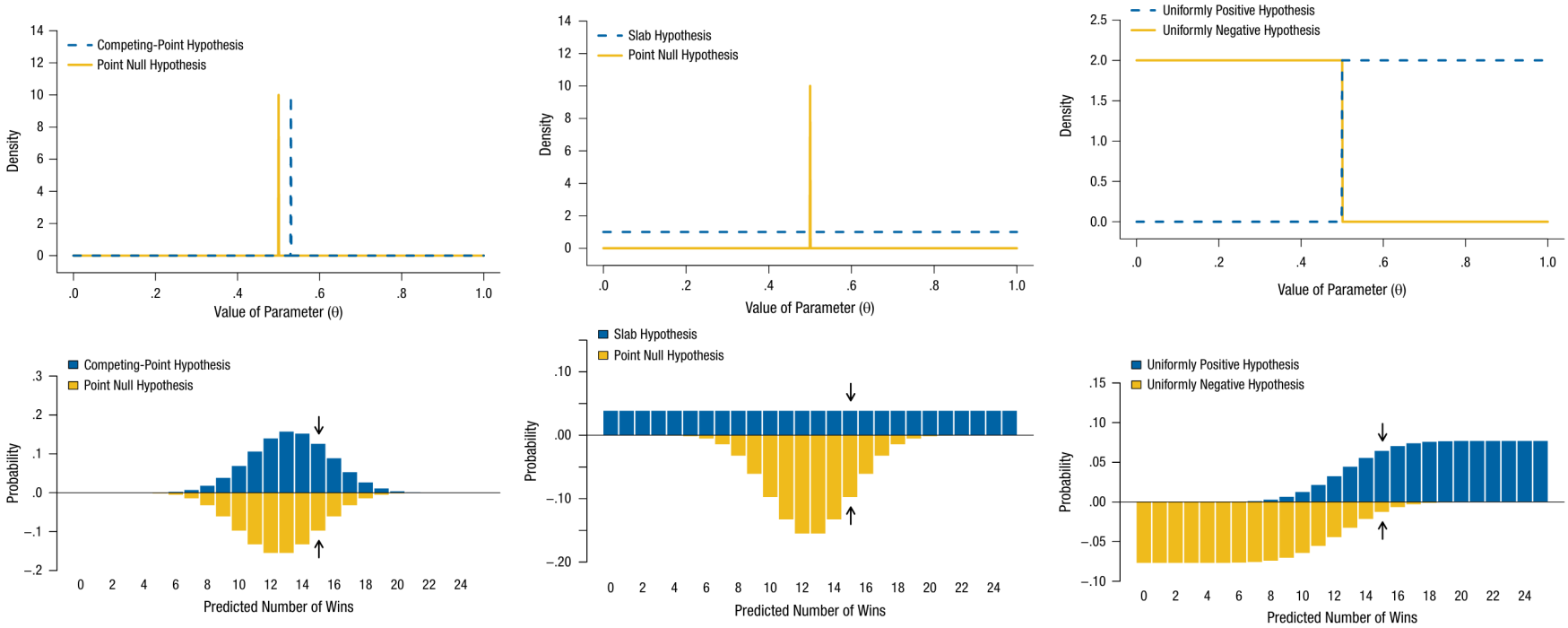
# Example: Point vs. range




(Etz et al., 2018).

- Point .5 vs. uniformly positive hypothesis for binary guessing task.
- *BF* is the ratio of the heights of the two bars for a given number of wins.

# Further examples




(Etz et al., 2018).

# Building & comparing models


- Creative process that requires actively thinking about the problem, difficult to automatise.
- But increasing tools to facilitate it, e.g. the bayesTestR package.

(Etz et al., 2018; Makowski et al., 2019).

# Sensitivity analysis

# Types of sensitivity analysis

- Assess to what extent the analysis results depend on prior assumptions and modelling decisions, e.g. about
  - prior parameter means
  - prior parameter variances
  - distribution family
  - number of samples, sampling algorithm

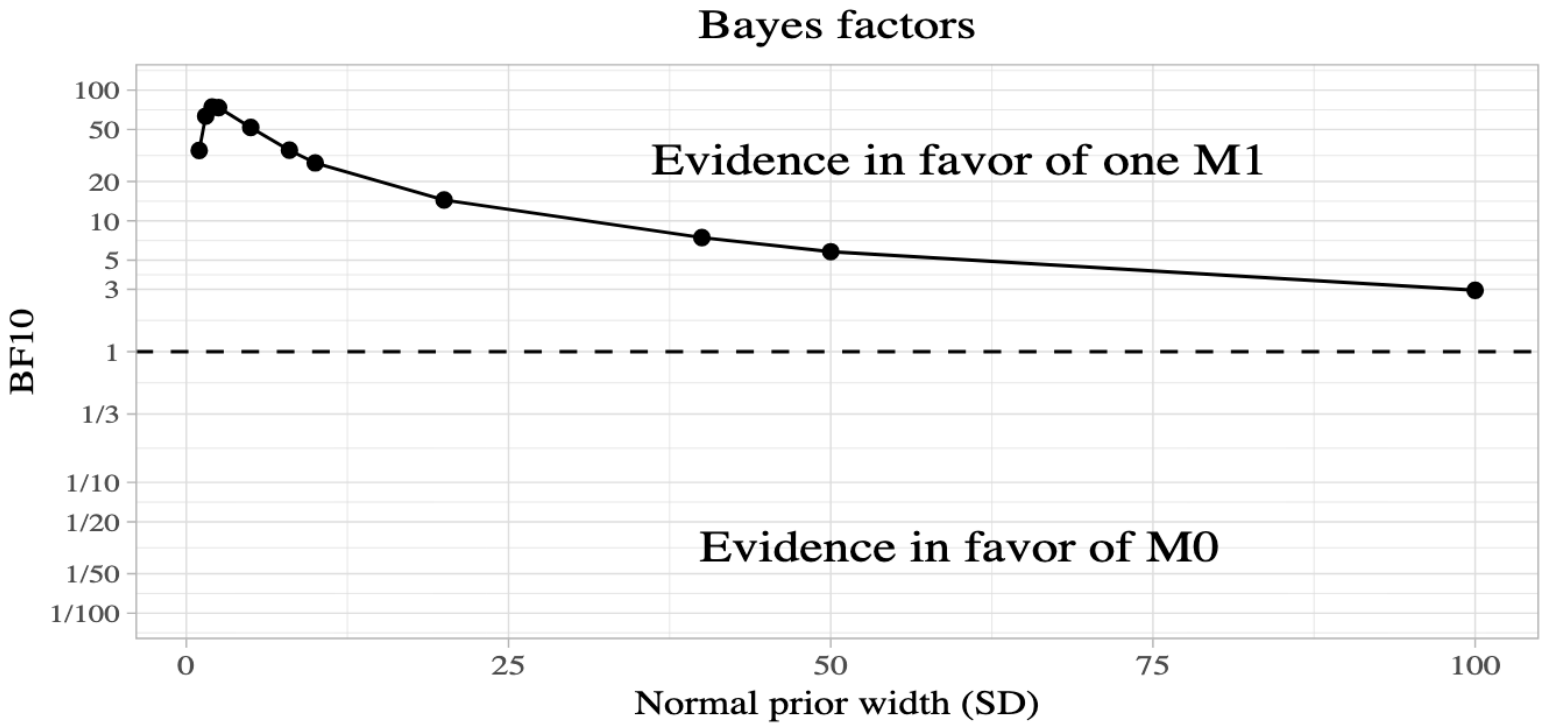
(Veenman et al., 2023).

# Types of sensitivity analysis

- Assess to what extent the analysis results depend on prior assumptions and modelling decisions, e.g. about
  - prior parameter means
  - prior parameter variances
  - distribution family
  - number of samples, sampling algorithm

(Veenman et al., 2023).

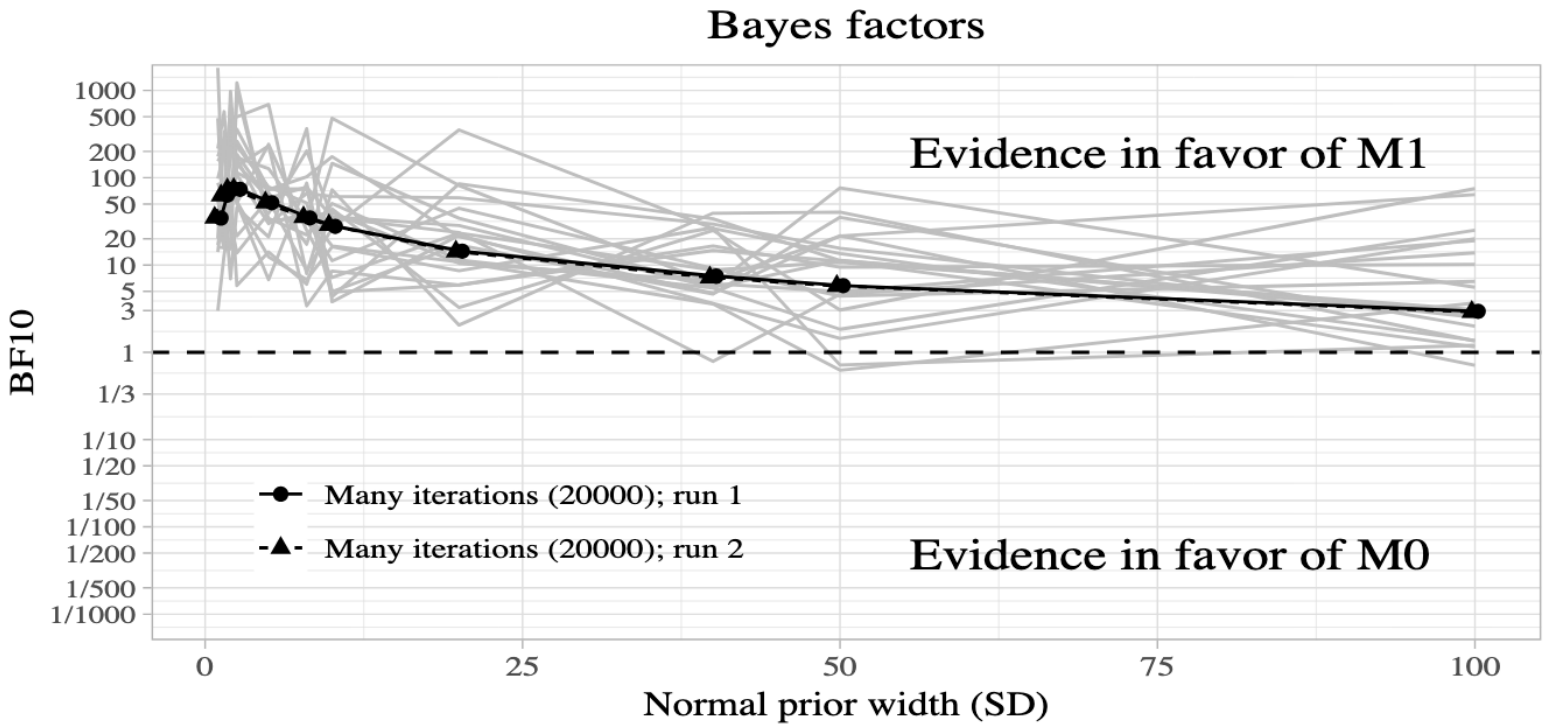
# *BF* sensitivity to prior *SD*




- Robust evidence for effect, but effect size probably small
- Wider priors on *SD* reduce evidence for effect

(Nicenboim, Schad, & Vasishth, 2025). *Note.* SD priors: [1, 1.5, 2, 2.5, 5, 8, 10, 20, 40, 50, 100].

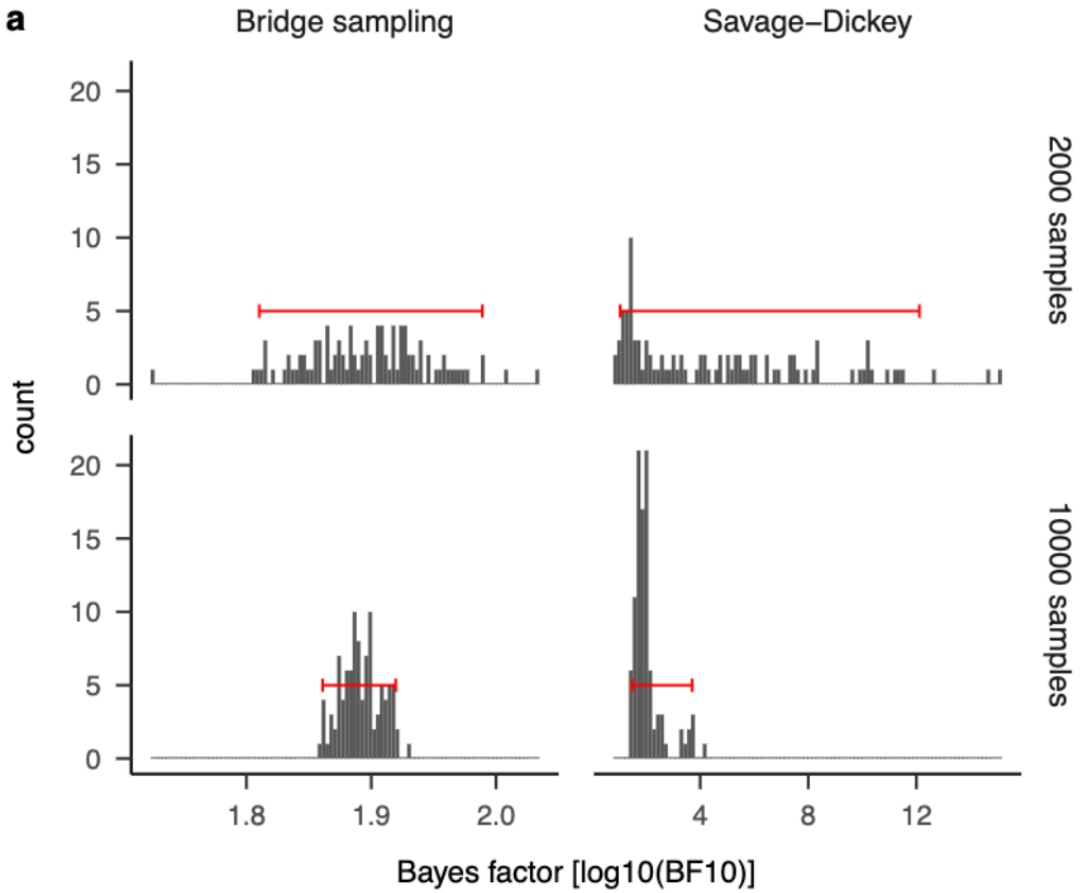
# *BF* sensitivity to iteration number




- Stable *BF* estimates require more samples than stable parameter estimates.

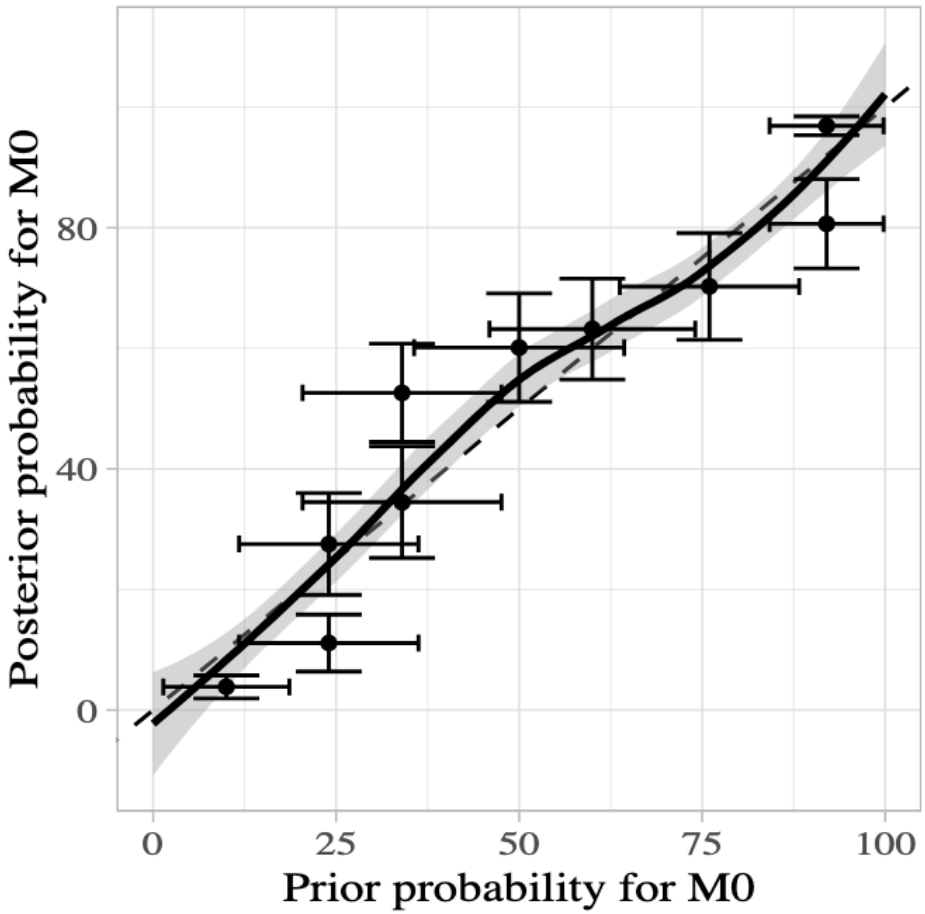
(Nicenboim, Schad, & Vasishth, 2025). *Note.* Black lines: 2 runs with 20 000 iterations. Gray lines: 20 runs with default number of iterations (2000).

# *BF* sensitivity to estimation method




(Schad et al., 2023). *Note.* Red horizontal error bars indicate 95 percent quantiles.

# Simulation based callibration




(Nicenboim, Schad, & Vasishth, 2025).

# Example with brms

# Penguin factors


```
> head(penguins)

  species    island bill_len bill_dep flipper_len body_mass
1  Adelie Torgersen     39.1     18.7          181      3750
2  Adelie Torgersen     39.5     17.4          186      3800
3  Adelie Torgersen     40.3     18.0          195      3250
```

Can a penguin's flipper length be predicted from its bill length? And from its bill depth?

# Models to compare

```r
m1 <- brm(flipper_len ~ 1
   , data = penguins
   , family = gaussian()
   , sample_prior = TRUE
   , warmup = 1000
   , iter = 10000
   , chains = 4
   , cores = 4
   , save_pars = save_pars(all = TRUE) # required to compute
     BFs
```

```r
m2 <- brm(flipper_len ~ bill_len)
```

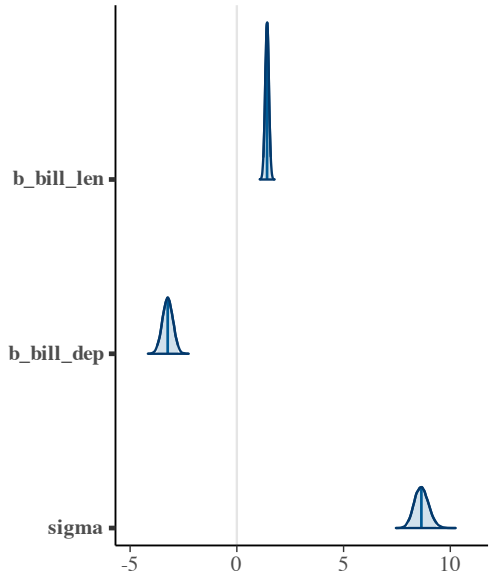
```r
m3 <- brm(flipper_len ~ bill_len + bill_dep)
```

# Absolute vs. relative evidence

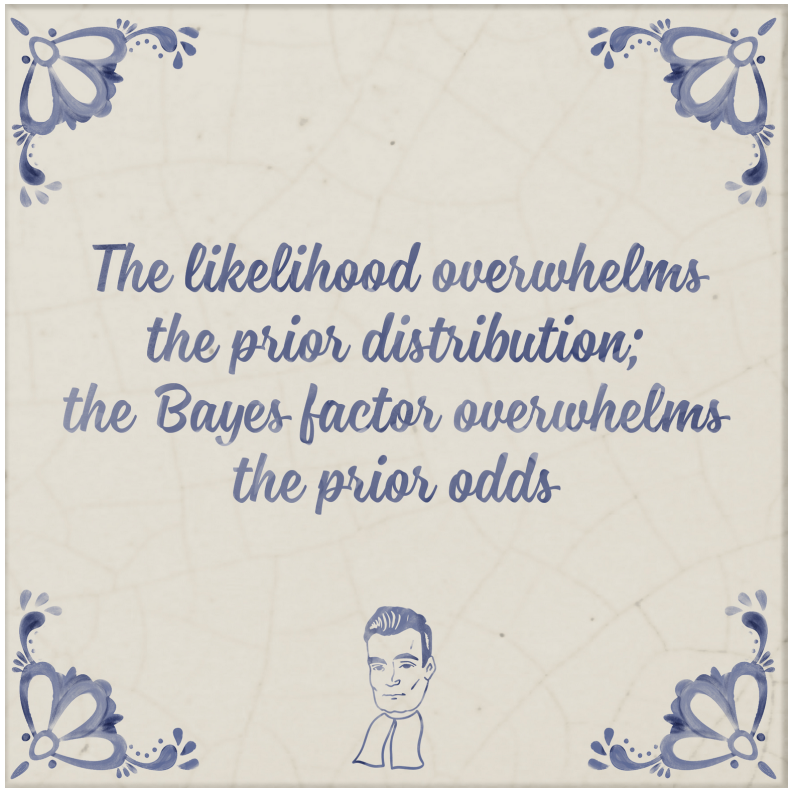
```
mar1 <- bridge_sampler(m1, silent = T)
mar2 <- bridge_sampler(m2, silent = T)
mar3 <- bridge_sampler(m3, silent = T)

bf_12 <- bayes_factor(mar1, mar2)
bf_23 <- bayes_factor(mar2, mar3)
bf_13 ?
```

$BF_{12} > 100$
$BF_{23} > 100$
$BF_{13}$?



# To be tested through sensitivity analysis




Source: www.bayesianspectacles.org