Bayesian statistics
3/4

Hypothesis testing

Oussama Abdoun (MEng, PhD) – oussama.abdoun@pm.me

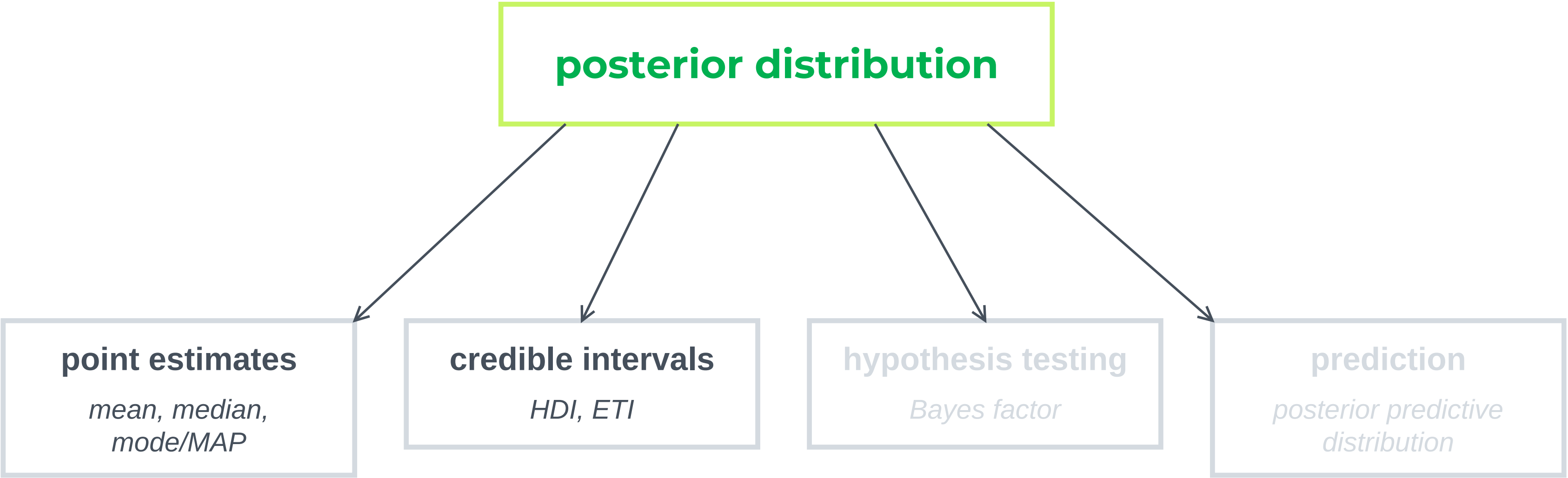
Bayesian Statistics – CRNL – dec 2024

1.

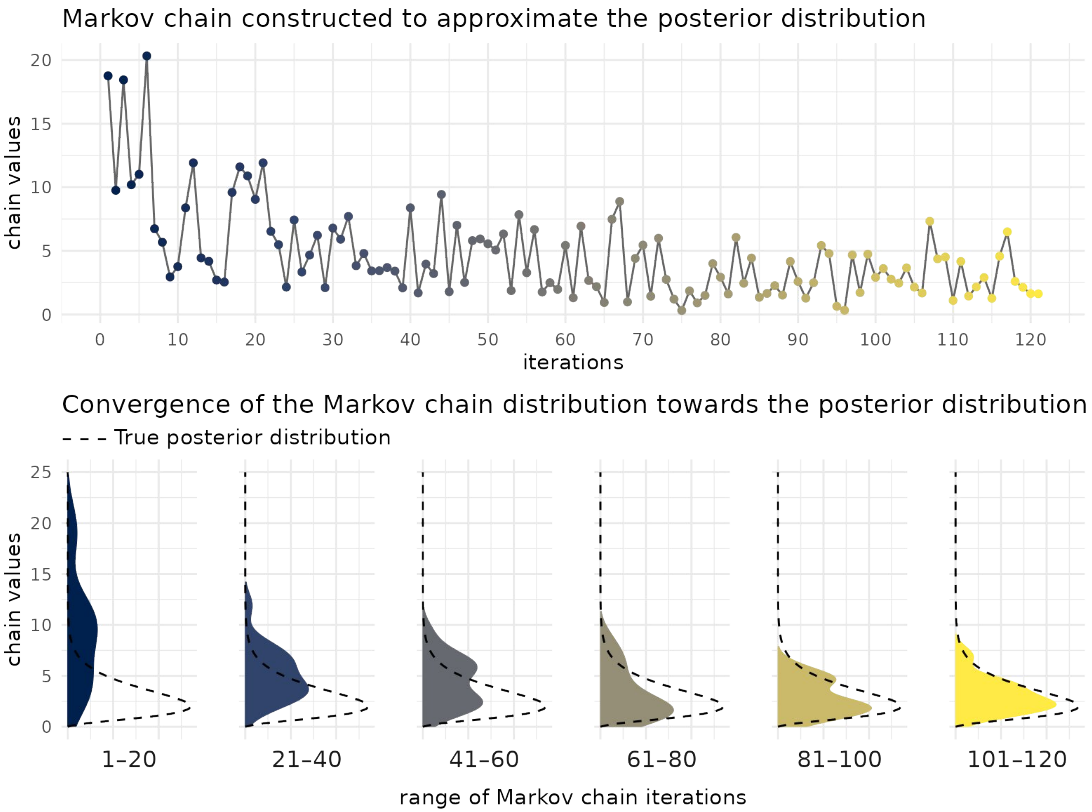
Posterior-based hypothesis testing

The central role of the posterior distribution

In Bayesian statistics, all results are derived from the posterior distribution



Numerical simulation of the posterior: MCMC



When prior distributions are **not conjugate distributions** of the likelihood, we don't have an explicit expression of the posterior distribution anymore and we need to calculate it **numerically**. We use **Markov chain Monte Carlo** (MCMC) techniques, a family of algorithms sharing the same basic procedure:

1. A **Markov chain** (= random process where each sample depends probabilistically on the previous one) is created such that it, *in the long run*, its distribution converges towards the true posterior distribution.

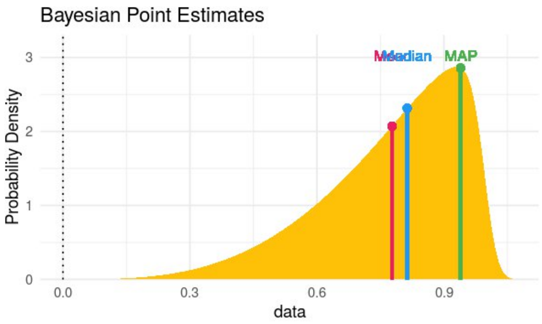
2. A large number of **samples** (several to tens of thousands) are generated **iteratively** from the Markov chain.

3. Initial samples (typically 1000) are considered as not converged yet and rejected ("**warm up**" phase); the rest of the samples is used as an **approximation of the posterior** distribution.

Point and interval estimates

`bayestestR` package in the `easystats` ecosystem

easystats.github.io/bayestestR/



point estimates

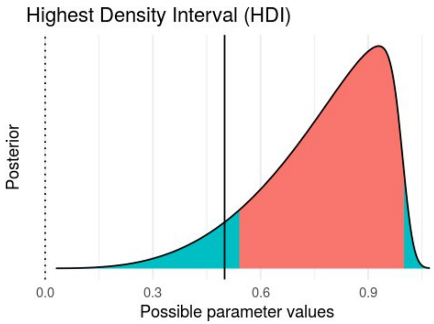
mean, median, mode/MAP

`point_estimate()`

credible intervals

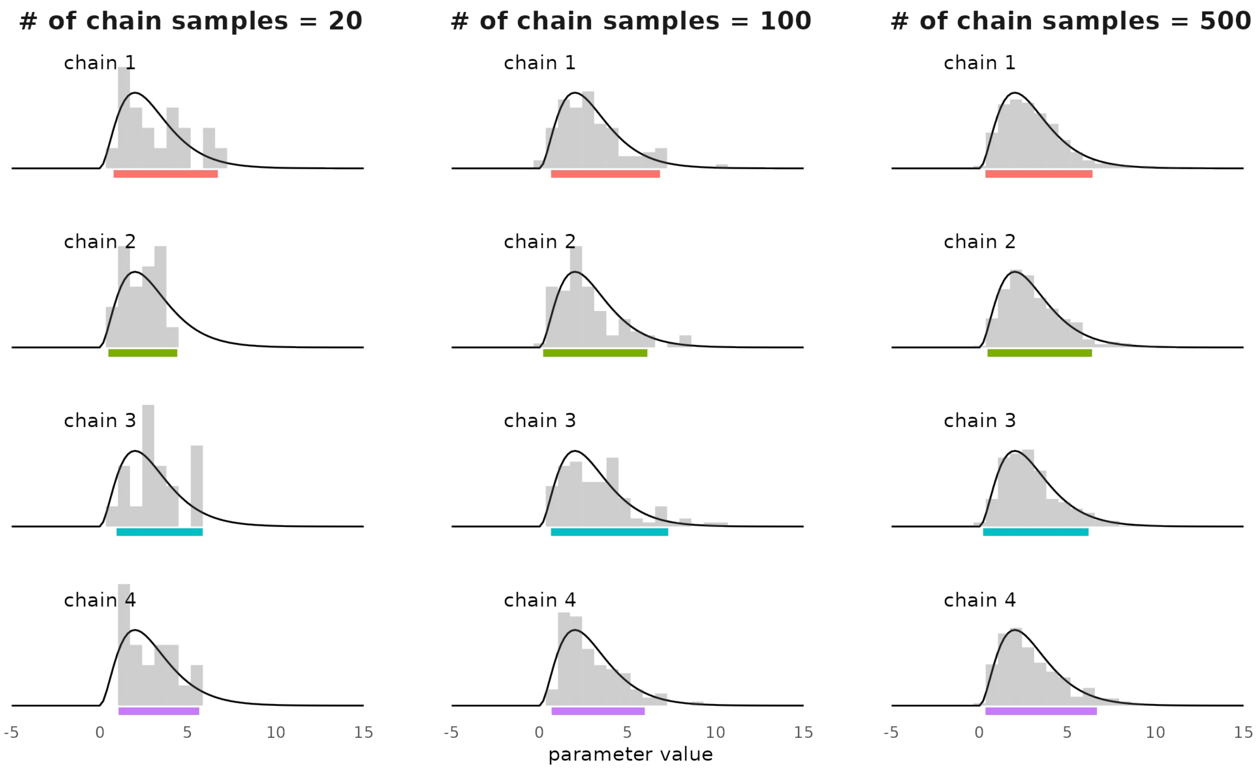
HDI, (ETI)

`hdi()`

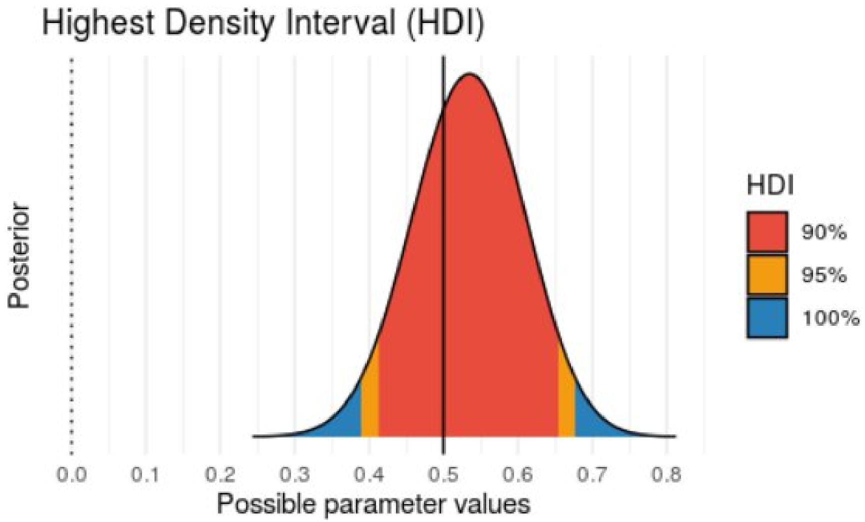


Credible intervals & numerical simulations

The more samples in the posterior distribution, **the more stable** the credible interval



Credible interval: 95% or 90%?



Compared to the 95%, the 90% credible interval is...

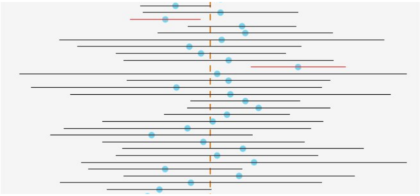
+ **more stable** to numerical errors
- **less conservative**

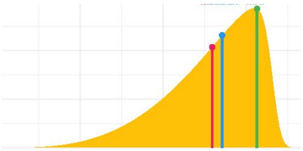
→ Use **95%** if there are more than **10.000 samples** of the posterior distribution

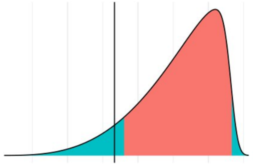
In `bayestestR`, the default is **89%** (!) to highlight the arbitrariness of the confidence level.

Frequentist vs. Bayesian statistics

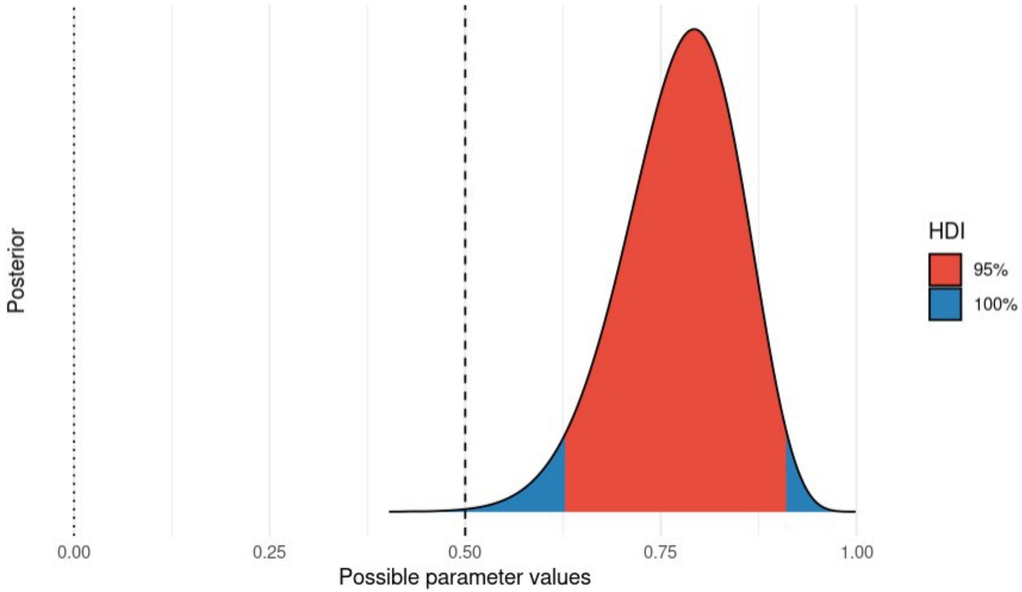
	Frequentist	Bayesian
Definition of probability	Long-run frequency of events	Degree of belief / certainty
View on model parameters	*True value:* unknown *Estimate:* fixed	*True value:* unknown *Estimate:* probabilistic
Method of estimation	From the data only	From the posterior (data + prior)
Uncertainty interval	"Confidence intervals" Confidence level (e.g. 95%) is a property of the procedure, not of the interval	"Credibility intervals" Confidence level (e.g. 95%) is a measure of the uncertainty around the estimate







Can we test a hypothesis on a parameter from its posterior distribution?



Hypothesis testing based on the posterior
Exact/point/precise hypothesis

The **credible interval** defines a whole **range of exact hypotheses** that can be **rejected** with high confidence.

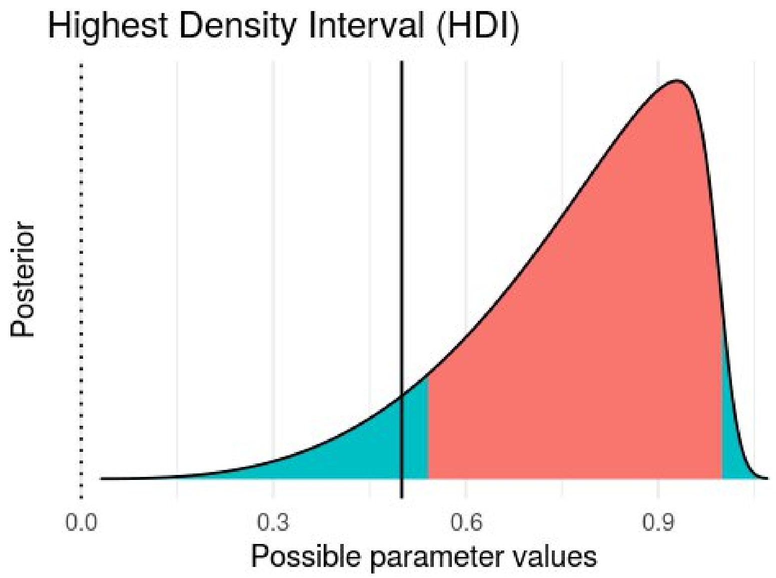
$$p\left(\theta_{low} \leq \theta \leq \theta_{high}\right) = .95$$

$$\Leftrightarrow p\left(\theta \notin [\theta_{low}, \theta_{high}]\right) = .05$$

$$\Longrightarrow p\left(\theta = \theta_0\right) < .05$$

Note: this is very different from the frequentist *p*-value which is $p\left(== y_{obs} \mid \theta = \theta_0\right)$

However, it does not allow to **accept** an exact hypothesis, only to reject it (at best).



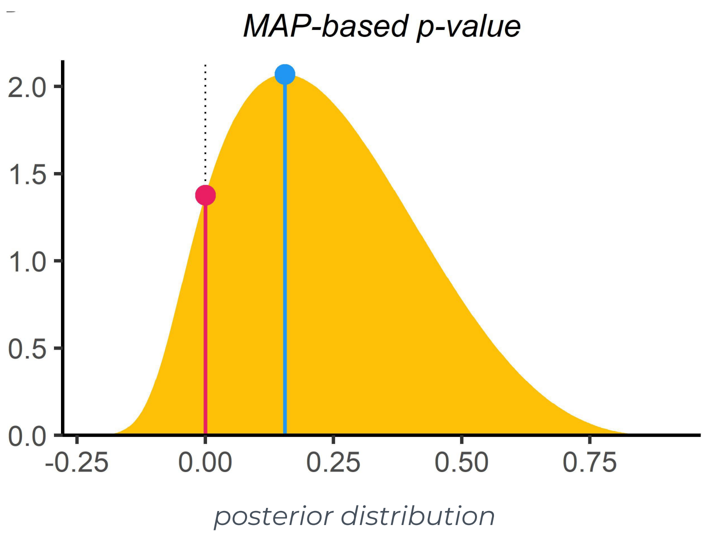
Hypothesis testing based on the posterior
Exact/point/precise hypothesis

For a continuous parameter, the **absolute probability** of an exact hypothesis, $p(H_0{:}\theta = \theta_0)$ is meaningless. That's why probability values are called **densities**.

But we could compare the probability of the null hypothesis with the probability of the most likely value (the MAP): $p(\theta = \theta_0)/p(\theta = \theta_{MAP})$

Limitations:
- ignores most of the information contained in the posterior distribution
- can not provide evidence *for* the null: at best, θ_0 is the MAP and p = 1

Strengths:
- no need for hypothesis-specific priors (unlike BF)
- no Jeffreys-Lindley-Bartlett paradox (unlike BF)



`p_pointnull()` in **bayestestR**

Mills, J. A. (2007). *Objective Bayesian Precise Hypothesis Testing.*

Hypothesis testing based on the posterior
Range hypothesis

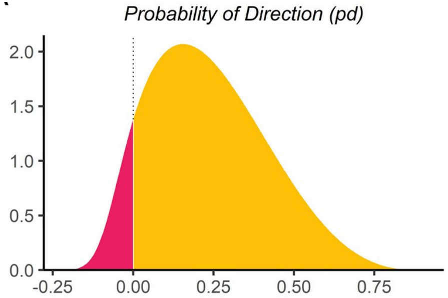
Arguments against point hypotheses

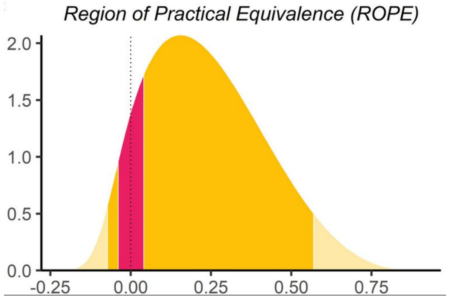
- **the null hypothesis can always be rejected:** trivial deviations due to measurement bias, sampling error and other uncontrolled factors will come out as statistically significant given sufficiently large sample sizes (Meehl 1978, Cohen 1994)
- rejecting the exact null hyothesis (the frequentist NHST approach) is a **weak form of theory testing**
- **theories rarely make exact hypotheses** (except when they specify a complete and detailed mechanism, e.g. physical equation).

Alternative: use a **range hypothesis** as a more powerful statement. Two possibilities:

- test a **direction hypothesis**
- test a null range, also called the **region of practical equivalence (ROPE)**, that encodes negligible effects

Range hypotheses can be rejected, accepted or neither (no conclusion).





`p_direction()` in **`bayestestR`**

`p_rope()` in **`bayestestR`**

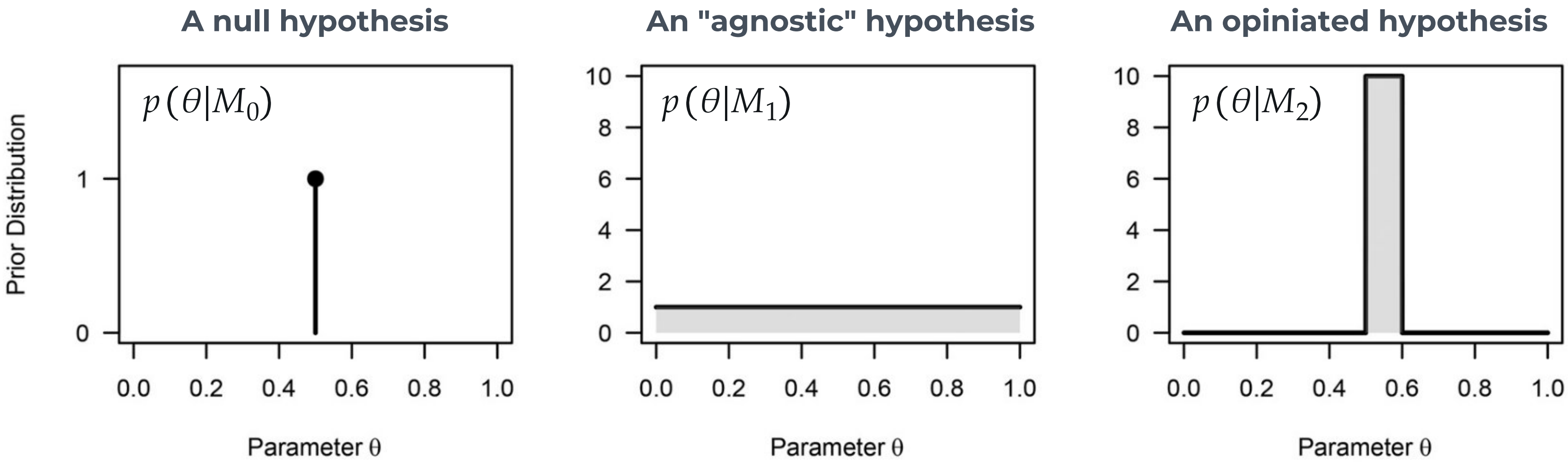
2.

Bayes Factors

Hypotheses as generative models

Different **hypotheses can be encoded as distinct models making specific predictions** for the data. Here, a model = statistical model + prior distribution for the parameters θ.

Examples for a binary outcome experiment (coin toss, medical treatment, test accuracy, etc.):



Bayes Factor

As a (marginal) likelihood ratio

Testing the likelihood of isolated hypotheses is of little interest. We are usually interested in how *competing* hypotheses are *differentially* supported by data.

Definition: the **Bayes Factor** is the ratio of the likelihoods of two statistical models, integrated over the prior probabilities of their parameters (“marginal likelihood”):

$$BF_{ij} = \frac{p(y|M_i)}{p\left(y|M_j\right)} =$$

Interpretation: “Model M_i predicts the data BF_{ij} times better than M_j”

If M_0 is a **null hypothesis**, then:

$$BF_{i0} = \frac{p(y|M_i)}{p(y|M_0)} = \frac{\int p(y|\theta_i)p(\theta_i)d\theta_i}{p(y|\boldsymbol{\theta}_0)}$$

Compare with the **likelihood ratio test**:

$$LR = \frac{p(y|\theta_{MLE})}{p(y|\theta_0)}$$

Bayes Factor
As relative belief updating

Applying the Bayes theorem:

$$BF_{10} = \frac{p(y|M_1)}{p(y|M_0)} = \frac{\dfrac{p(M_1|y)}{p(M_1)}}{\dfrac{p(M_0|y)}{p(M_0)}}$$

= belief updating of model M_1

= belief updating of model M_0

Thus, the Bayes factor is…

— …a continuous measure of evidence

— …a predictive updating factor

— …independent of models' prior probability

— …equal to the posterior model odds if models are equally probable *a priori*

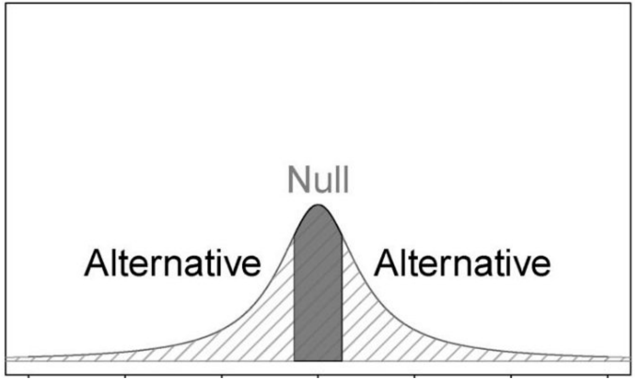
Which we can rewrite :

$$\underbrace{\frac{p(M_1|y)}{p(M_0|y)}}_{\textit{posterior model odds}} = \underbrace{BF_{10}}_{\textit{likelihood ratio}} \times \underbrace{\frac{p(M_1)}{p(M_0)}}_{\textit{prior model odds}}$$

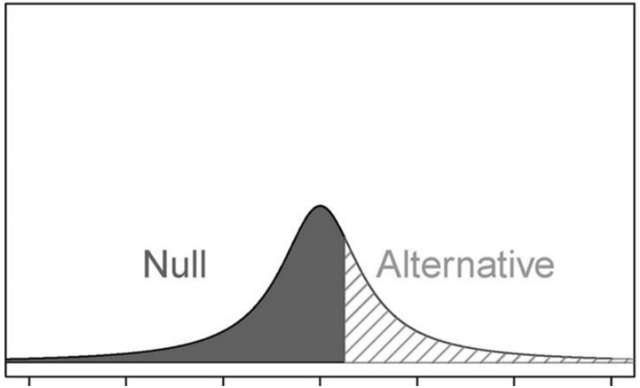
Initially, I believed M_1 was times more likely than M_0. After seeing the new data, which is times better predicted by M_1 than M_0, I now believe M_1 is times more likely than M_0. Therefore, my prior belief ratio has changed by a factor

Bayes Factor for range hypotheses

The calculation of the BF is straightforward when the competing models are **non-overlapping, complementary intervals from the same distribution**



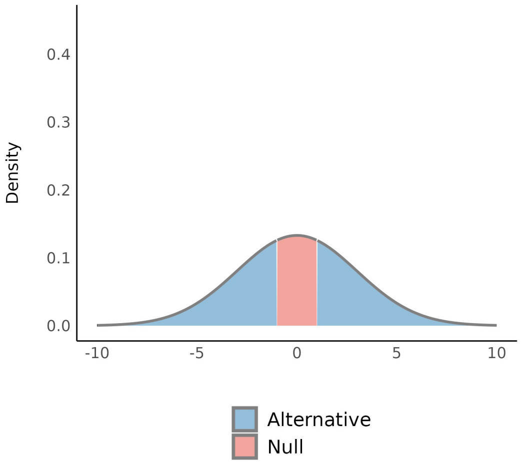
“two-sided”



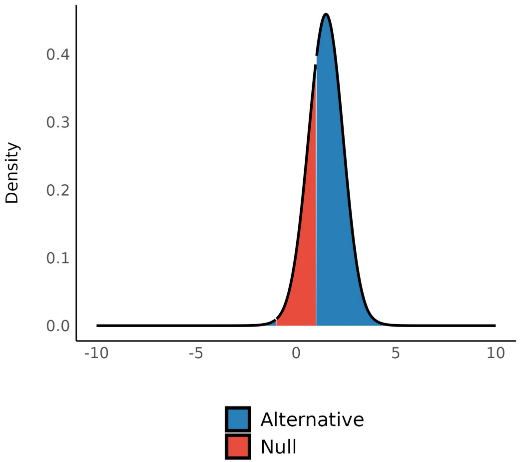
“one-sided”

Bayes Factor for range hypotheses

prior distribution

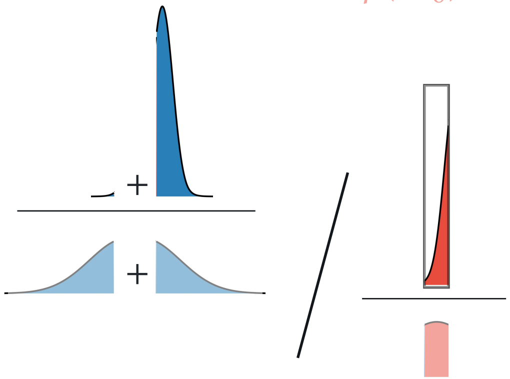


posterior distribution



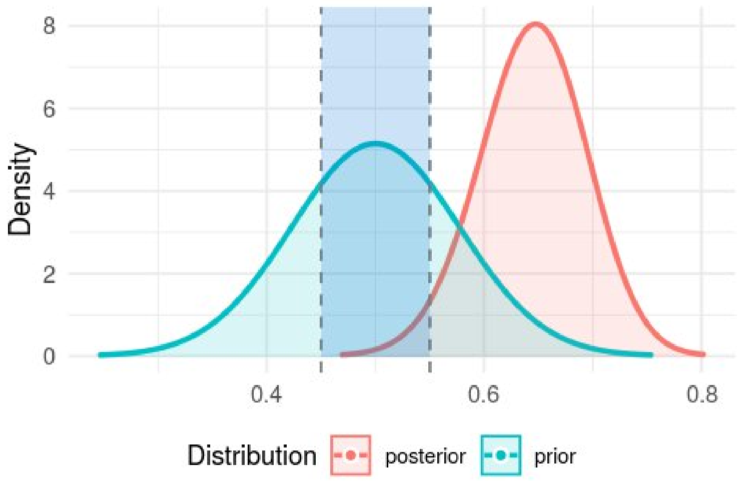
Bayes Factor

$$BF_{10} = \frac{p(M_1|y)}{p(M_1)} \Big/ \frac{p(M_0|y)}{p(M_0)}$$

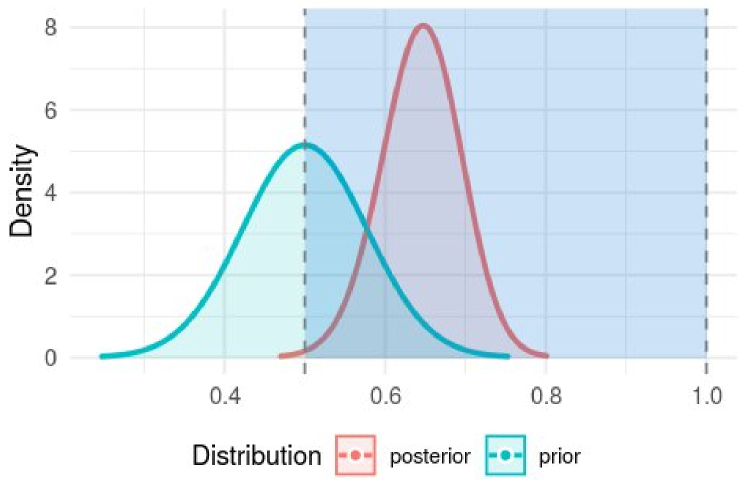


Bayes Factor for range hypotheses

Region of practical equivalence (ROPE)



Probability of direction



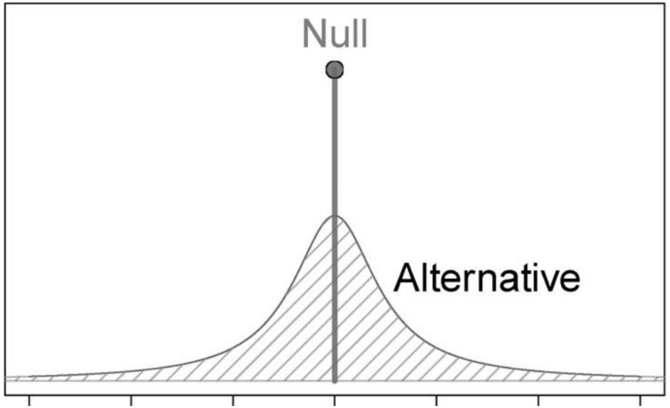
`bf_params()` in **bayestestR**

Bayes Factor for exact hypotheses

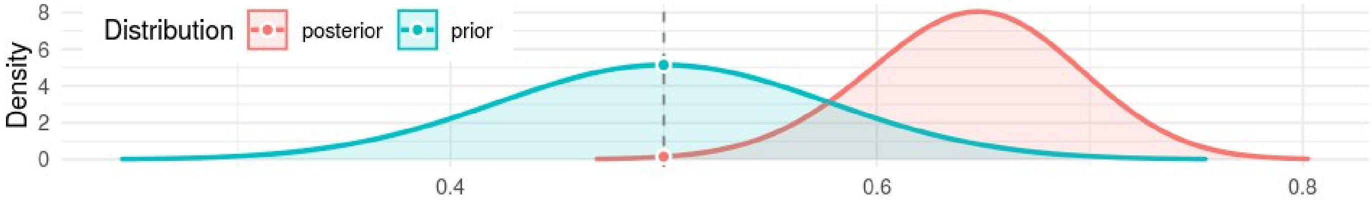
Savage-Dickey density ratio

Let H_0 be an exact null hypothesis ($H_0: \theta = \theta_0$) and H_1 the complementary hypothesis ($H_1: \theta \neq \theta_0$). Then:

$$BF_{01} = \frac{p(\theta = \theta_0 | y)}{p(\theta = \theta_0)}$$



This special case of the Bayes Factor is called the **Savage-Dickey density ratio**.



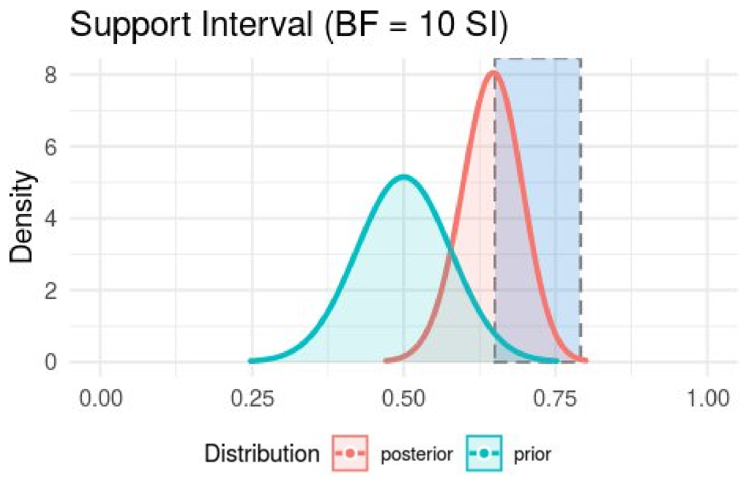
`bf_params()` in `bayestestR`

Bayes Factor for exact hypotheses

Support interval

Which values of the parameter are best supported by data?

Support interval = all values for which the Savage-Dickey density ratio is above a certain threshold (here, 10).



`si()` in **`bayestestR`**

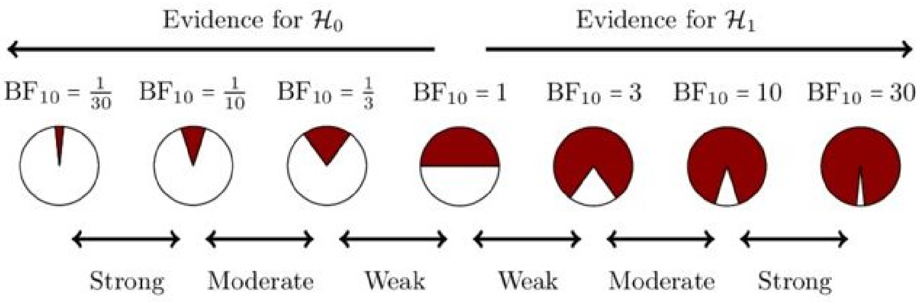
Bayes Factor

Measure of evidence

Conventional interpretation of Bayes factor values (Kass & Raftery 1995)

BF	$log_{10}BF$	Strength of evidence
1 to 3	0 to 1/2	Barely worth mentioning
3 to 10	1/2 to 1	Substantial
10 to 100	1 to 2	Strong
> 100	> 2	Decisive

JASP



Don't replace the p-value dichotomous ritual by a BF multichotomous ritual!!

Bayes Factor

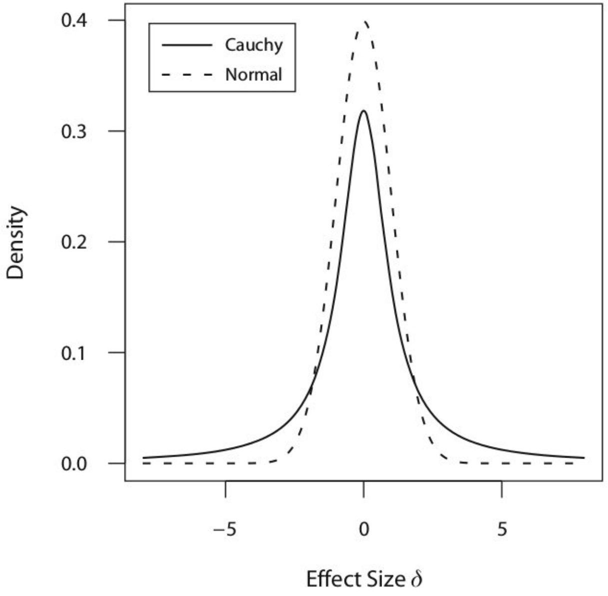
Application to the two-sample location test

Approach implemented in **BayesFactor** and **JASP**:

— parametrize the model in terms of the standardized effect size (~ Cohen's d): $\delta = {}^{\mu}/_{\sigma}$

— prior = Cauchy distribution with scale r (~ variance)
⇒ fatter tails than the normal distribution

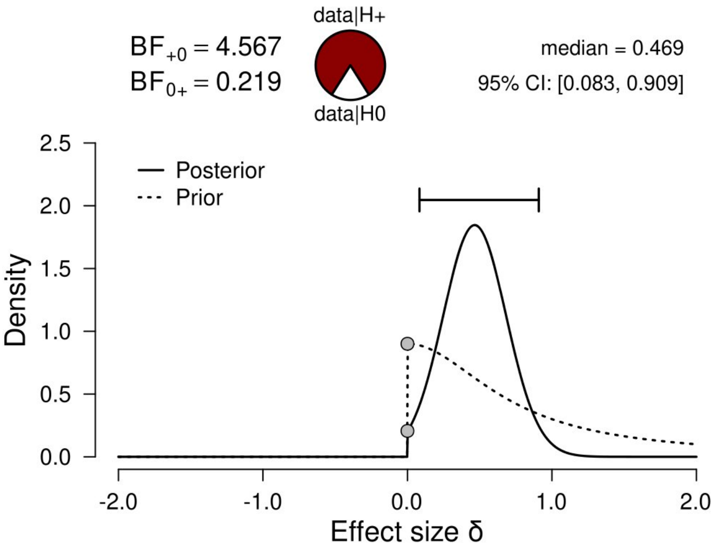
`ttestBF()` in the `BayesFactor` package



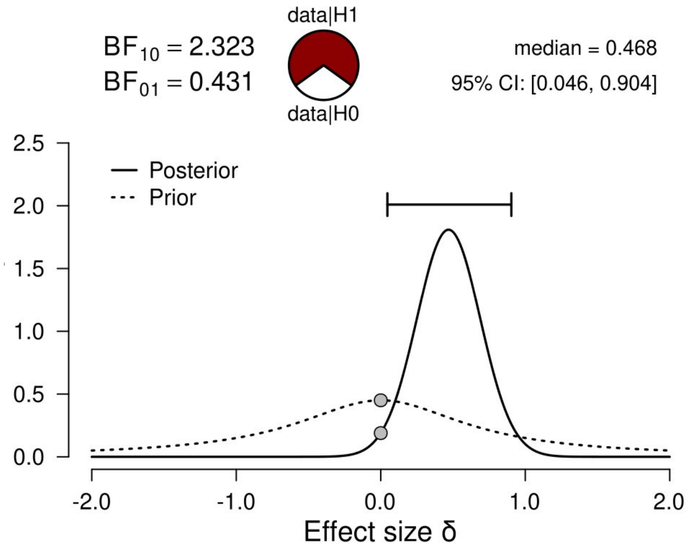
Rouder, Speckman, Sun, Morey & Iverson (2009). *Bayesian t tests for accepting and rejecting the null hypothesis.* Psychonomic Bulletin & Review doi.org/10.3758/PBR.16.2.225

Bayes Factor

Application to the two-sample location test



(a) One-sided analysis for testing: $H_+ : \delta > 0$

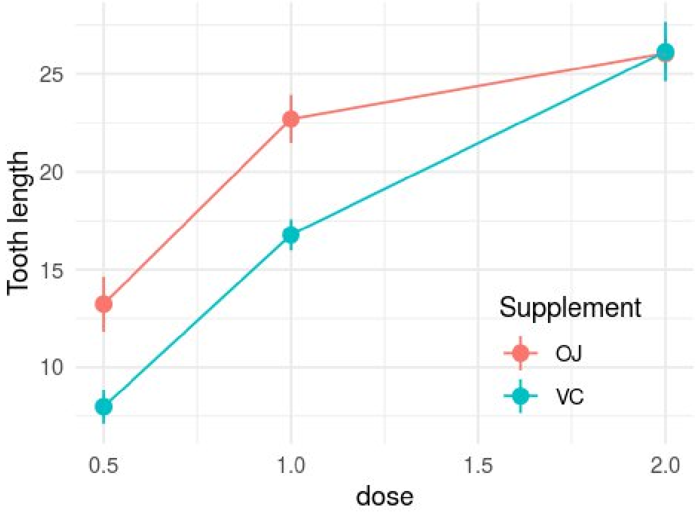


(b) Two-sided analysis for estimation: $\mathcal{H}_1 : \delta \sim \text{Cauchy}$

Bayes Factor
Application to the ANOVA

Guinea pigs are assigned to one of two treatments (vitamin C or orange juice) in one of three doses. The effect on tooth growth is measured.

or



Bayes Factor
Application to the ANOVA

New challenges:

— multiple variables ⇒ multiple parameters

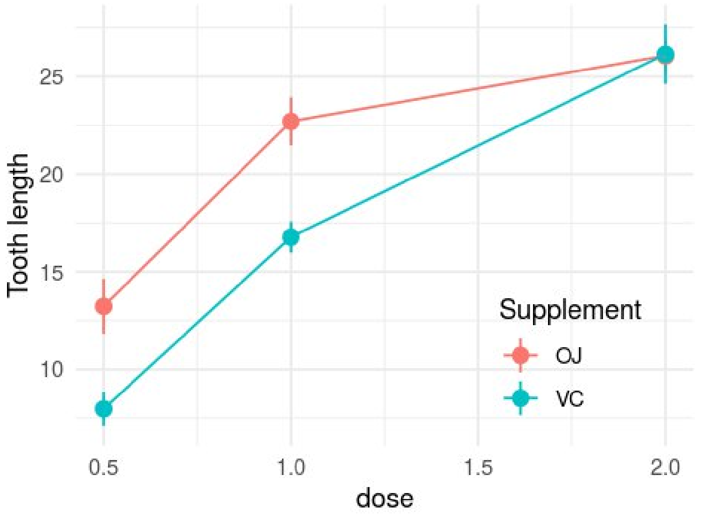
— a single variable (here, *dose*) can be encoded with **2** parameters

Solution implemented in *BayesFactor* and *JASP*

= **comparison between nested models**
e.g. with (M_2) and without (M_1) the interaction

$$BF_{21} = \frac{p\,(y|M_2)}{p\,(y|M_1)}$$

BF apply not only to models with *different priors*, but also to models with *different structures*!



Bayes Factor
Comparing multiple models

Transitivity

Imagine we have k models to compare: M_1, M_2, ..., M_k

To compare them, we don't need to calculate all pairwise BF. Thanks to the **transitivity** of Bayes Factors, we only need BF,**transitivity**

$$BF_{32} = \frac{p(M_3|y)}{p(M_2|y)} = \frac{p(M_3|y)}{p(M_1|y)} \times \frac{p(M_1|y)}{p(M_2|y)} = \frac{BF_{31}}{BF_{21}}$$

Posterior model probability

Using the definition of Bayes factors and the Bayes law, we can show that:

$$p(M_i|y) = \frac{BF_{i1}.p(M_i)}{\sum_{j=0}^{k} BF_{j1} \times p(M_j)}$$

which can be simplified, **when the prior distribution is uniform** over the model space:

$$p(M_i|y) = \frac{BF_{i1}}{\sum_{j=0}^{k} BF_{j1}}$$

Bayes Factor
Application to multiple regression

Suppose we study a phenomenon for which we have several candidate predictors.
For example, the concentration of a pollutant in the air might depend on temperature, humidity and/or wind:

```
conc ~ temp, humid, wind
```

Instead, we can try all possible models, with any combination of the 3 predictors.
Each can be included or not, so there are $2^3 = 8$ possible models:

null		**1 predictor**		**2 predictors**		**3 predictors**	
`~ 1`	M_0	`~ temp`	M_1	`~ temp, humid`	M_4	`~ temp, humid, wind`	M_7
		`~ humid`	M_2	`~ temp, wind`	M_5		
		`~ wind`	M_3	`~ humid, wind`	M_6		

see also JASP blog post

Bayesian Model Averaging (BMA)

Application to multiple regression

Now we can calculate the posterior probability of including `temp` as a predictor by summing up the posterior probabilities of the models where it shows up:

$$p(incl_{temp}|y) = p(M_1|y) + p(M_4|y) + p(M_5|y) + p(M_7|y)$$

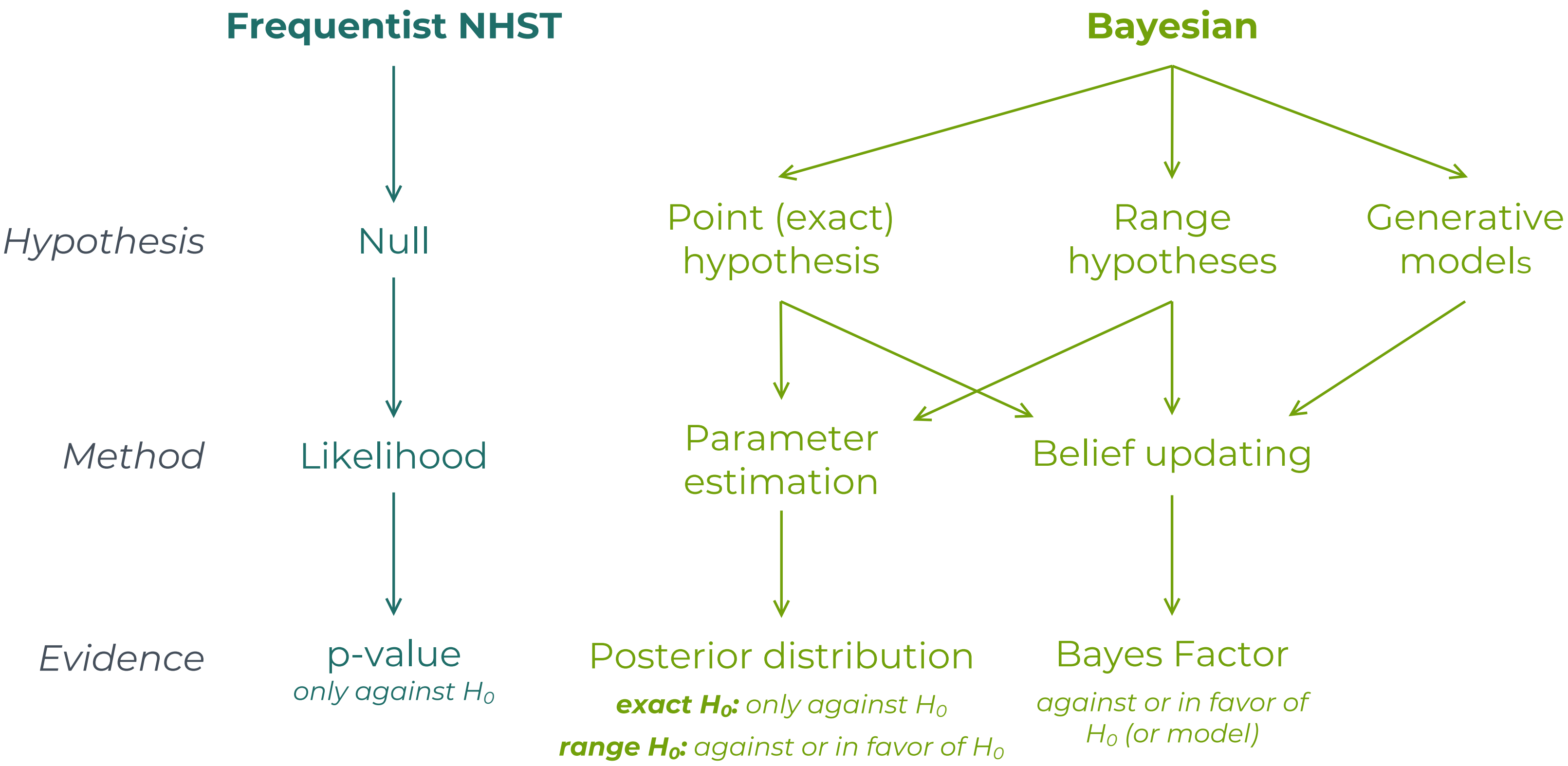
We can run a similar calculation for the *prior* probability of including `temp` as a predictor and derive a **Bayes factor of inclusion**: $BF_{incl} = p(incl_{temp}|y) \,/\, p(incl_{temp})$

The **estimation** of the effect of a specific predictor (e.g. temperature `temp`) depends on the model considered. Instead of relying on a single model, we average on all models that include `temp` as a predictor, weighted by their respective posterior model probabilities to obtain a **weighted posterior distribution** of the regression coefficient for `temp`.

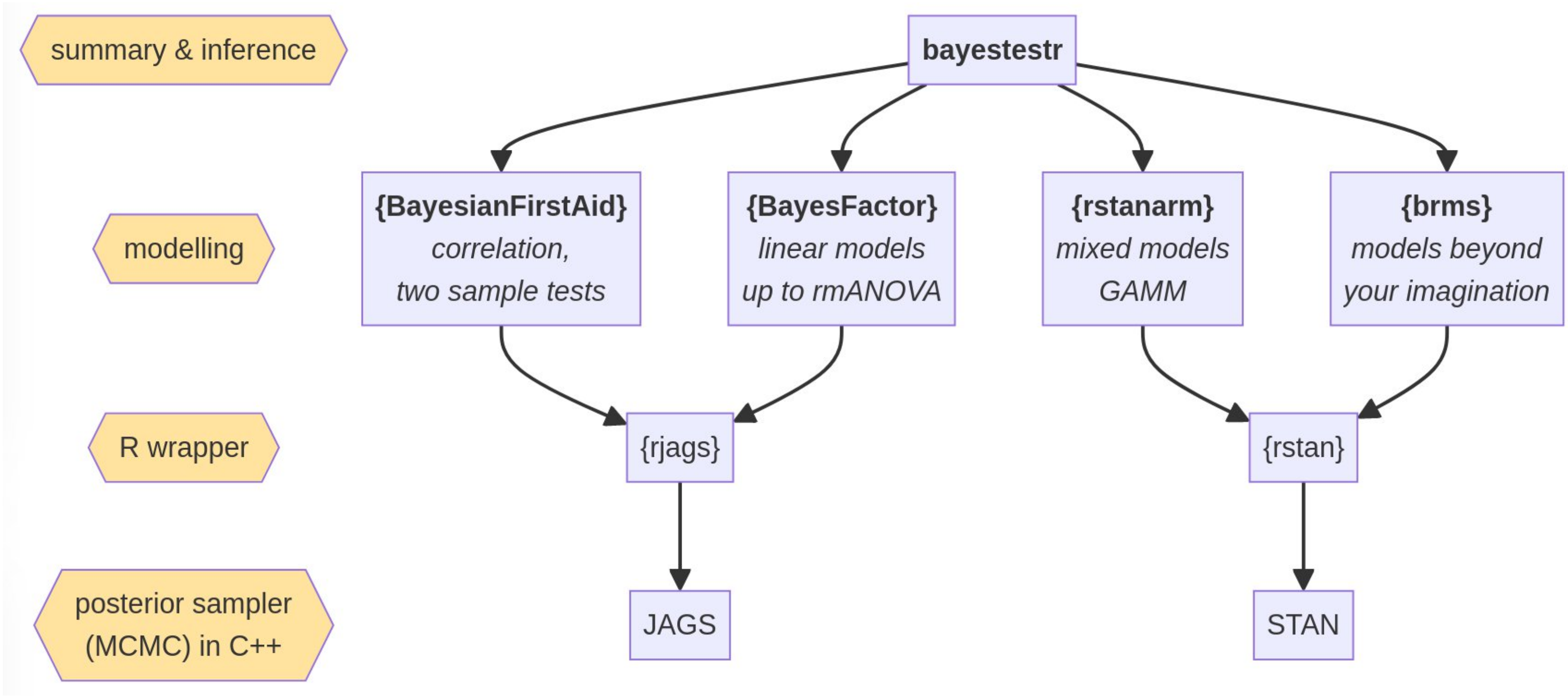
null		1 predictor		2 predictors		3 predictors	
`~ 1`	M_0	`~ temp`	M_1	`~ temp, humid`	M_4	`~ temp, humid, wind`	M_7
		`~ humid`	M_2	`~ temp, wind`	M_5		
		`~ wind`	M_3	`~ humid, wind`	M_6		

$$p(\beta_{temp}|y) = p(\beta_{temp}\ M_1) \times p(M_1|y) + \ldots + p(\beta_{temp}|y,M_7) \times p(M_7|y)$$

The many ways of Bayesian hypothesis testing



Software & package ecosystem



Frequentist vs. Bayesian statistics

	Frequentist	Bayesian
Definition of probability	Long-run frequency of events	Degree of belief / certainty
View on model parameters	Fixed	Probabilistic
Point estimates	Derived from the sample	Derived from the posterior distribution
Interval estimates	Confidence interval ; confidence level is a property of the procedure, not of the intervals themselves	Credibility intervals ; confidence level is a statement about the uncertainty of the model parameters
Hypothesis testing	Point hypotheses only Can only reject a hypothesis	Point and range hypotheses Can select the best one among multiple
Limitations	**Interpretability** **Usefulness**	**Time consuming (prior + computation)** **Lack of standards, rapid evolution**