Bayesian statistics 3/4

Hypothesis testing

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

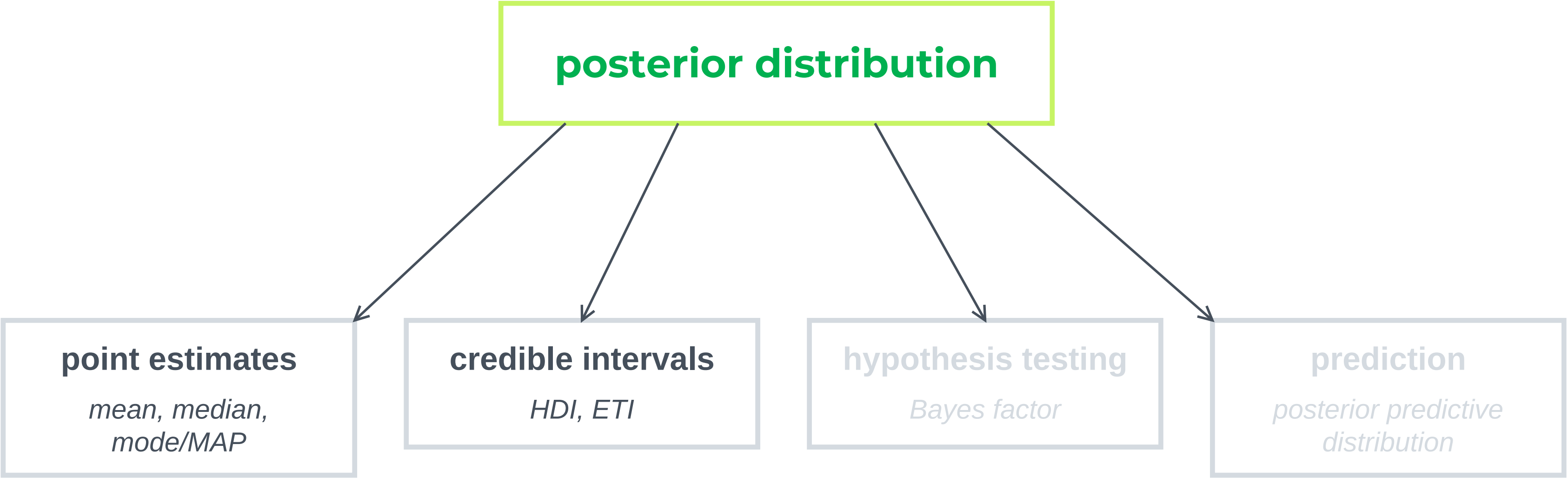
Bayesian Statistics – CRNL – dec 2024

1.

Posterior-based hypothesis testing

The central role of the posterior distribution

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



posterior distribution

point estimates

mean, median, mode/MAP

credible intervals

HDI, ETI

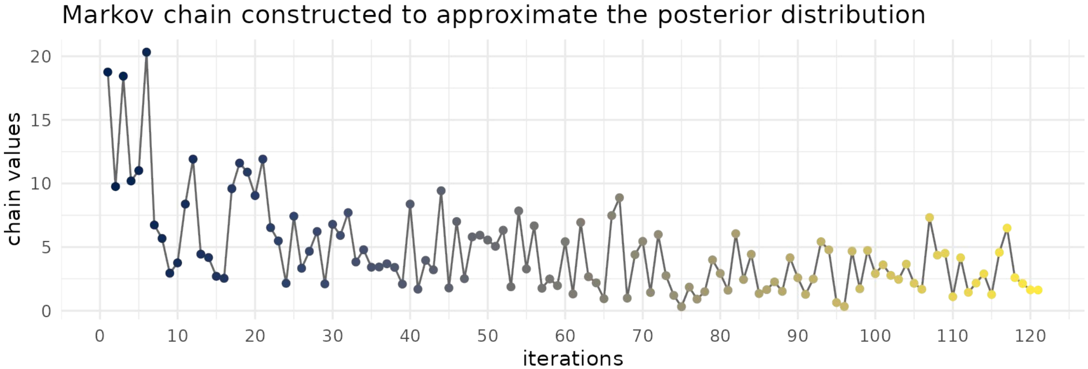
hypothesis testing

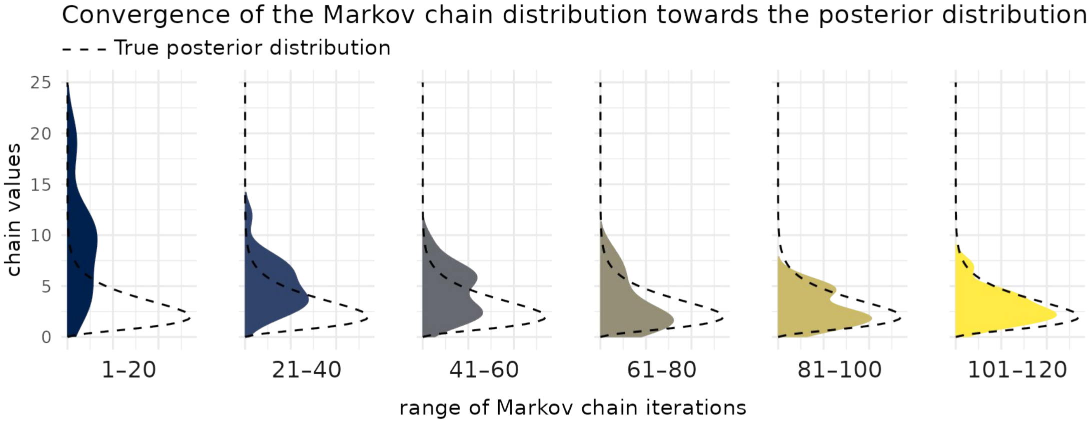
Bayes factor

prediction

posterior predictive 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.

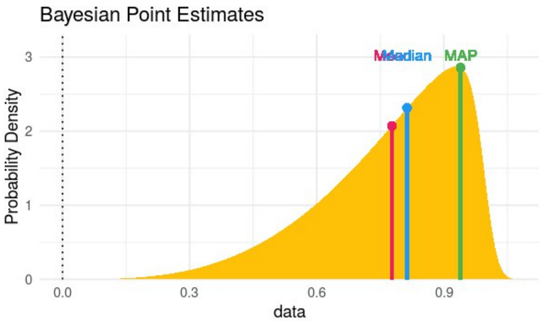
2. A large number of **samples** (several to tens of thousands) are generated **iteratively** from the Markov chain.

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

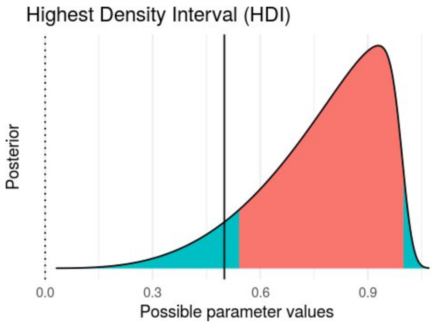
mean, median, mode/MAP

`point_estimate()`

credible intervals

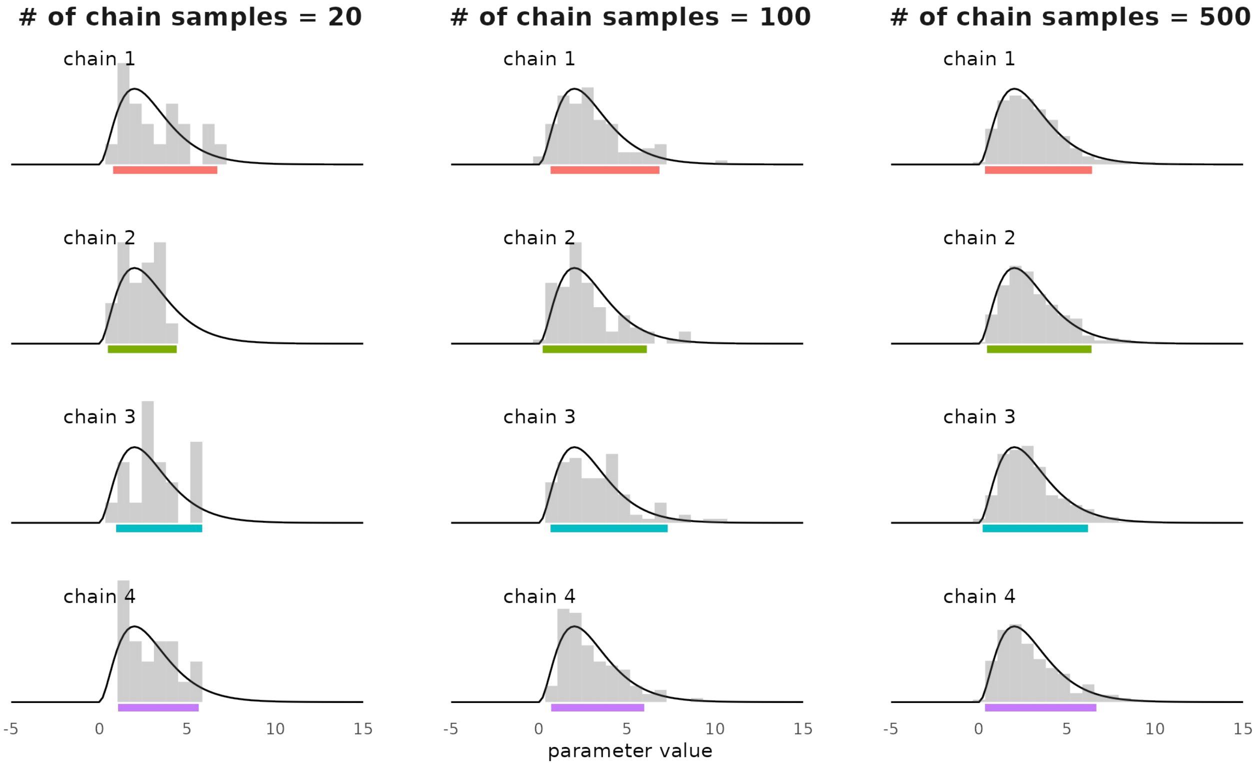
HDI, (ETI)

`hdi()`

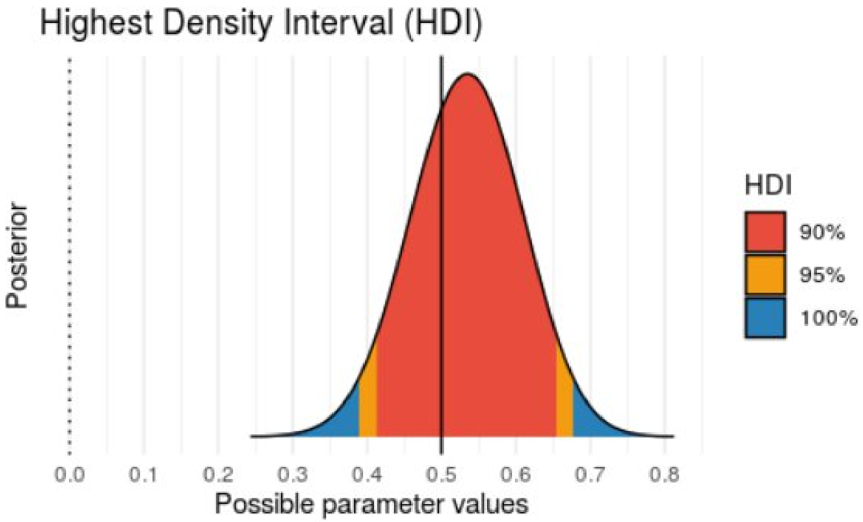


Credible intervals & numerical simulations

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…

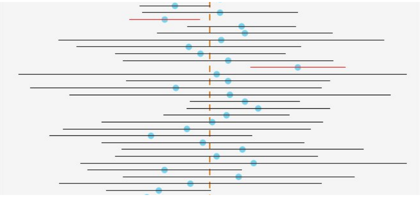
+ more stable to numerical errors
– less conservative

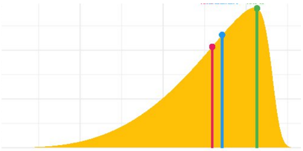
→ Use **95%** if there are more than **10.000 samples** of the posterior distribution

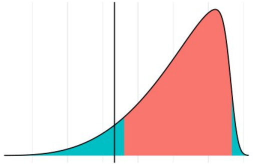
In `bayestestR`, the default is **89%** (!) to highlight the arbitrariness of the confidence level.

Frequentist vs. Bayesian statistics

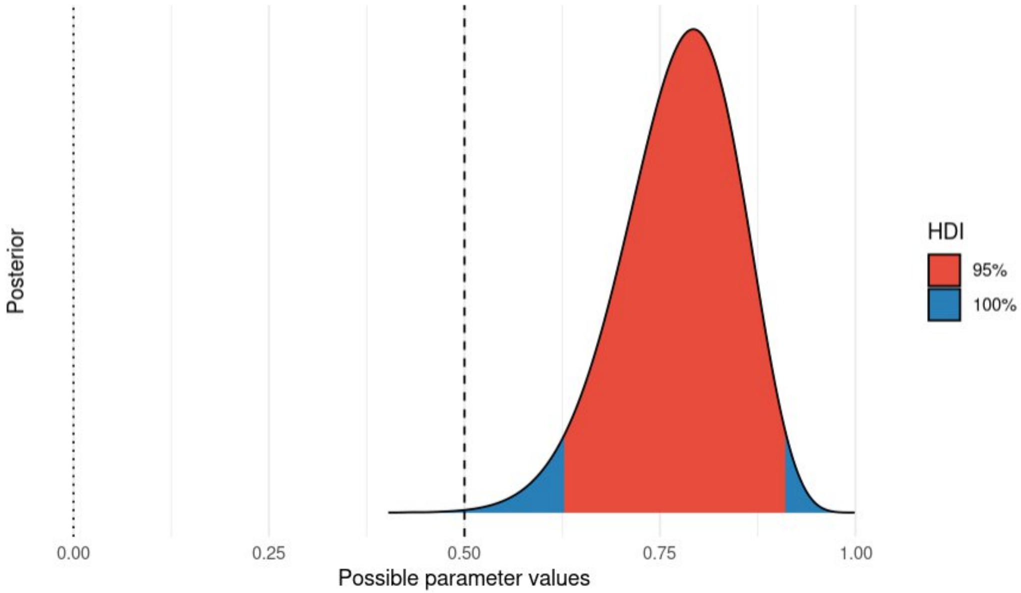
	Frequentist	Bayesian
Definition of probability	Long-run frequency of events	Degree of belief / certainty
View on model parameters	*True value*: unknown *Estimate*: fixed	*True value*: unknown *Estimate*: probabilistic
Method of estimation	From the data only	From the posterior (data + prior)
Uncertainty 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.

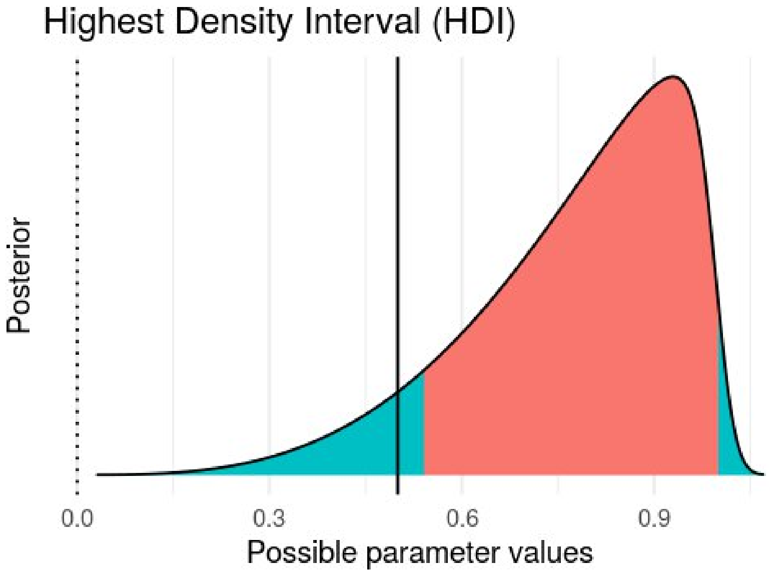
$$p\left(\theta_{low} \leq \theta \leq \theta_{high}\right) = .95$$

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

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

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, $p(H_0 \colon \theta = \theta_0)$ is meaningless. That's why probability values are called **densities**.

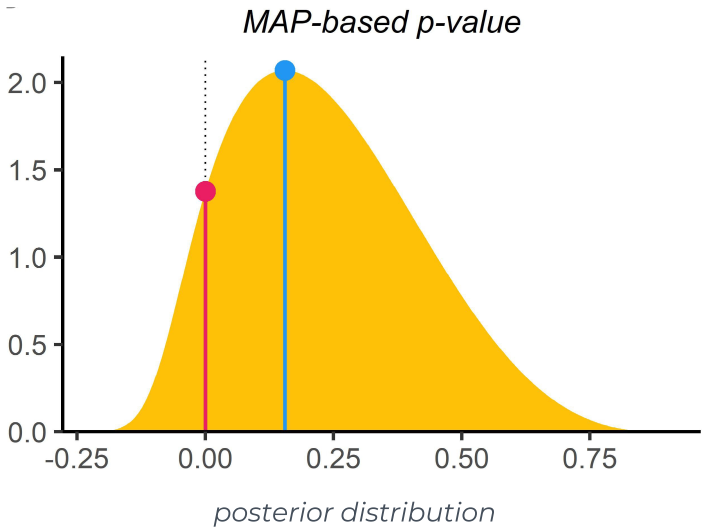
But we could compare the probability of the null hypothesis with the probability of the most likely value (the MAP): $p(\theta = \theta_0) / p(\theta = \theta_{MAP})$

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

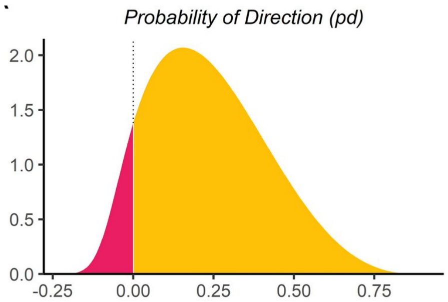
- **the null hypothesis can always be rejected:** trivial deviations due to measurement bias, sampling error and other uncontrolled factors will come out as statistically significant given sufficiently large sample sizes (Meehl 1978, Cohen 1994)
- rejecting the exact null hyothesis (the frequentist NHST 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).

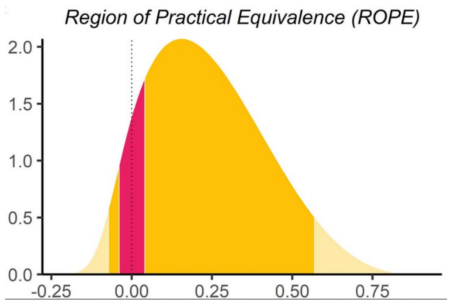
Alternative: use a **range hypothesis** as a more powerful statement. Two possibilities:

▣ 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).


Probability of Direction (pd)


Region of Practical Equivalence (ROPE)

`p_direction()` in `bayestestR`

`p_rope()` in `bayestestR`

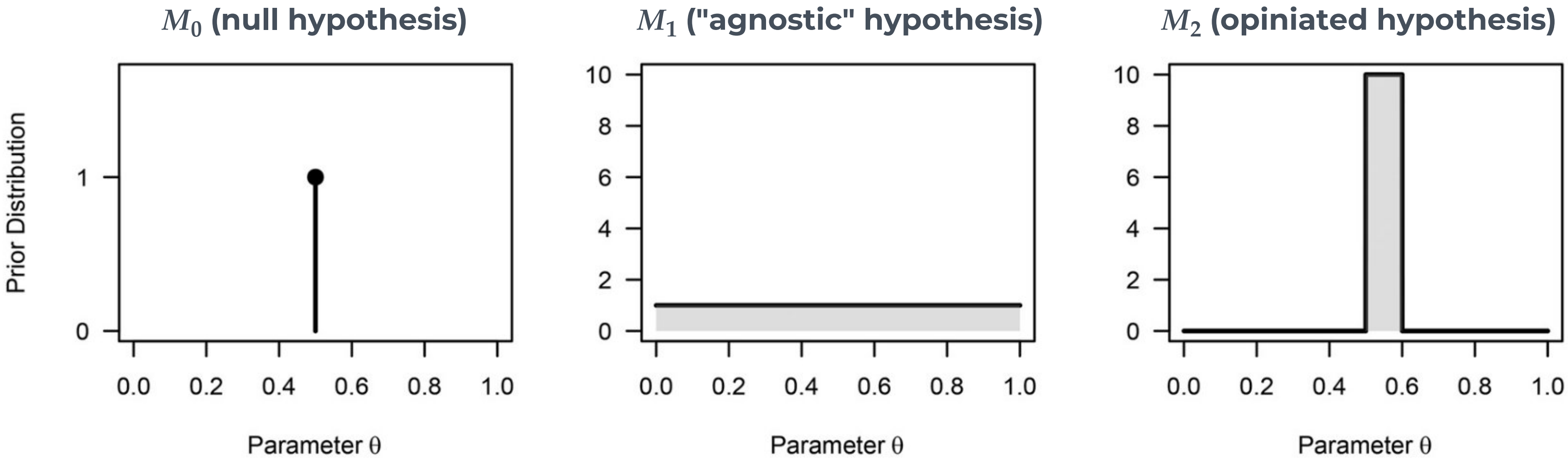
2.

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.):



M_0 **(null hypothesis)**

M_1 **("agnostic" hypothesis)**

M_2 **(opiniated hypothesis)**

Bayes Factor

As a (marginal) likelihood ratio

Definition: 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):

$$BF_{21} = \frac{p(y|M_2)}{p(y|M_1)} = \int \frac{\int p(y|\theta_2)p(1\theta_2)d\theta_2}{\int p(y|\theta_1)p(\theta_1)d\theta_1}$$

Interpretation: “Model M_2 supports the data BF_{21} times more than M_1”

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

$$BF_{10} = \frac{p(y|M_1)}{p(y|M_0)} = \frac{\int p(y|\theta_1)p(\theta_1)d\theta_1}{p(y|\boldsymbol{\theta}_0)}$$

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:

$$BF_{10} = \frac{p\,(y|M1)}{p\,(y|M_0)} = \frac{\dfrac{p\,(M_1|y)}{p(M_1)}}{\dfrac{p\,(M_0|y)}{p(M_0)}}$$

= belief updating
of model M_1

= belief updating
of model M_0

Which we can rewrite :

$$\frac{p\,(M_2|y)}{p\,(M_0|y)} = BF_{10} \times \frac{p(M_1)}{p(M_0)}$$

posterior
model odds

prior
model odds

Thus, the Bayes factor is…

— …a continuous measure of evidence

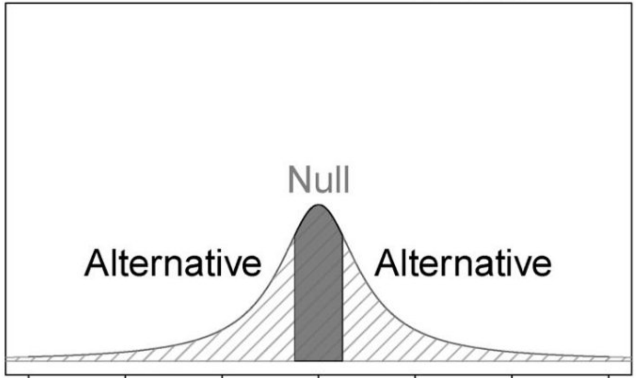
— …a predictive updating factor

— …independent of models' prior probability

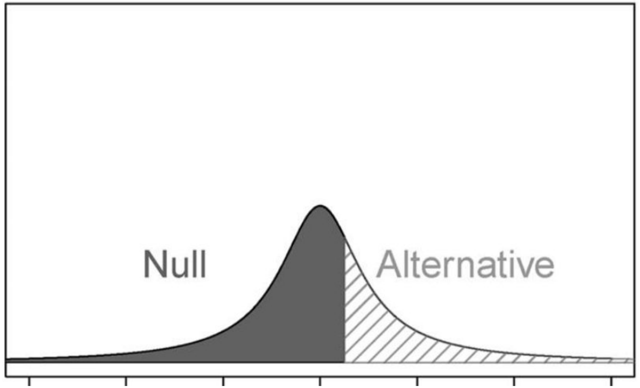
— …equal to the posterior model odds if models are equally probable *a priori*

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**



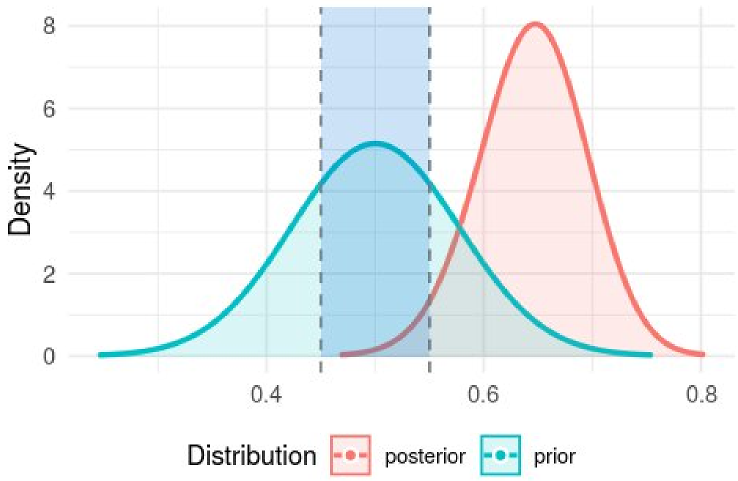
“two-sided”



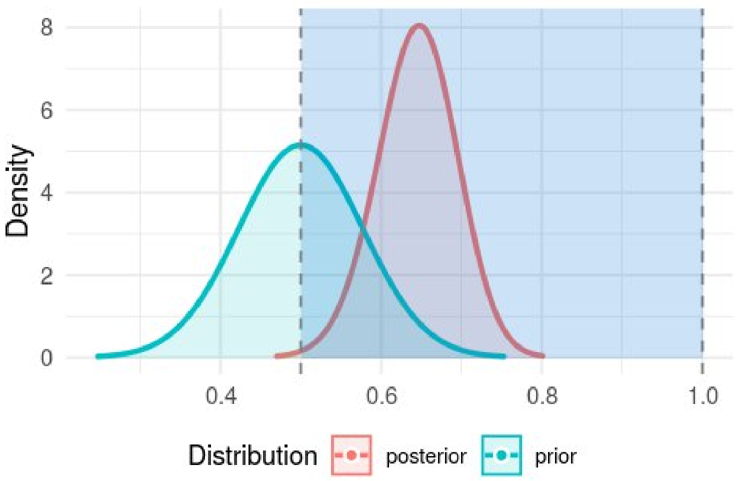
“one-sided”

Bayes Factor for range hypotheses

Region of practical equivalence (ROPE)



Probability of direction



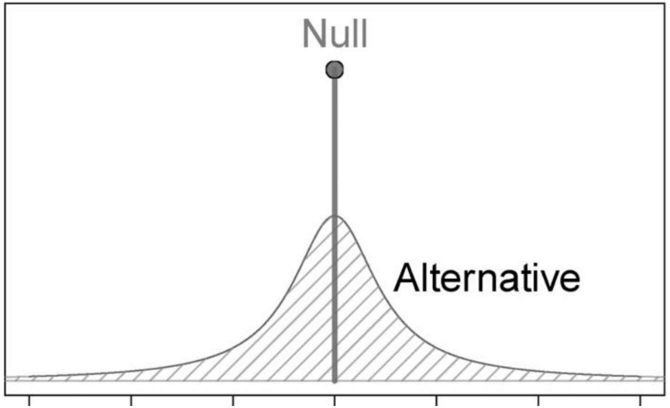
`bf_params()` in **bayestestR**

Bayes Factor for exact hypotheses

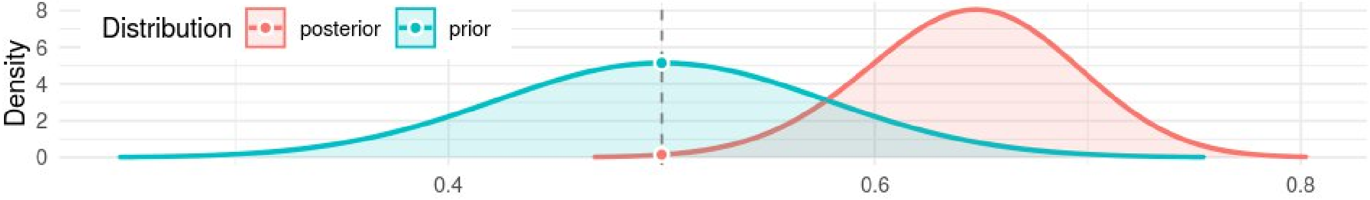
Savage-Dickey density ratio

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:

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



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



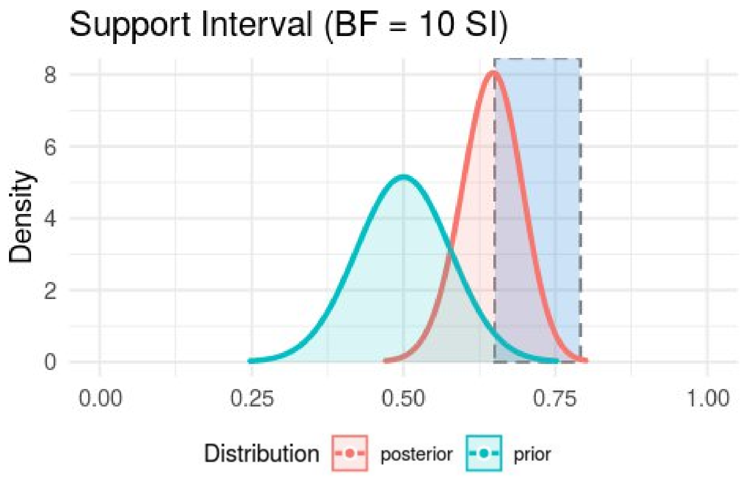
`bf_params()` in **`bayestestR`**

Bayes Factor for exact hypotheses

Support interval

Which values of the parameter are supported by data?

Support interval = all values for which the Savage-Dickey density ratio is above a certain threshold



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

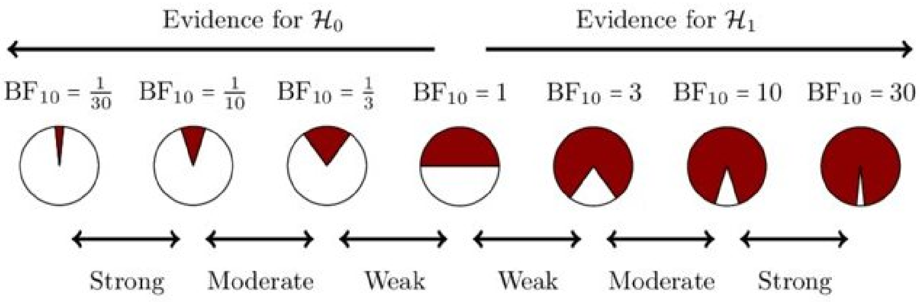
Bayes Factor

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!!**

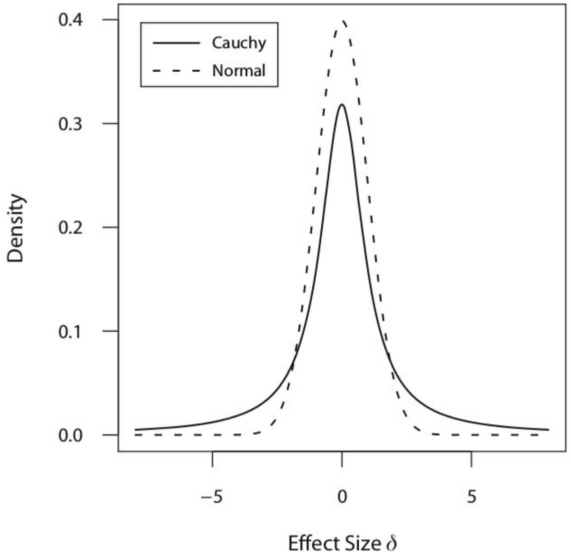
Bayes Factor

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

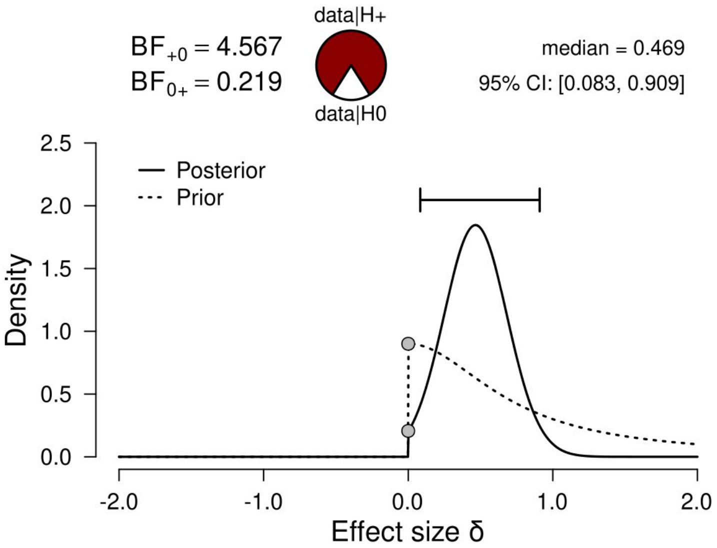


`ttestBF()` in the `BayesFactor` package

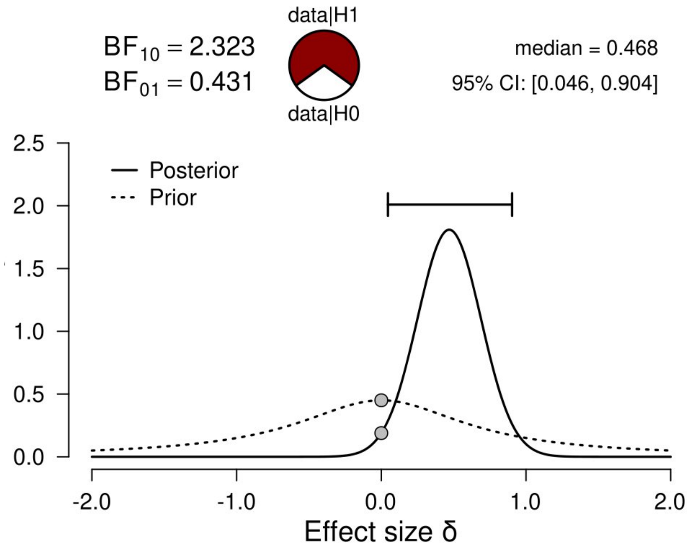
Rouder, Speckman, Sun, Morey & Iverson (2009). *Bayesian t tests for accepting and rejecting the null hypothesis.* Psychonomic Bulletin & Review doi.org/10.3758/PBR.16.2.225

Bayes Factor

Application to the two-sample location test



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



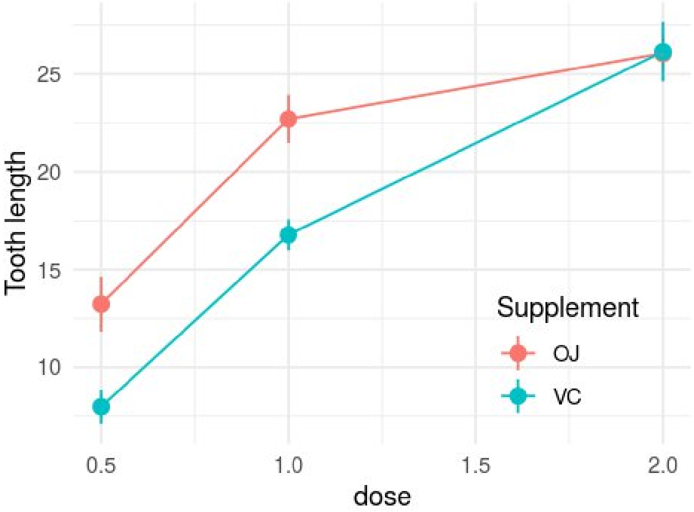
(b) Two-sided analysis for estimation:
$\mathcal{H}_1 : \delta \sim \text{Cauchy}$

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

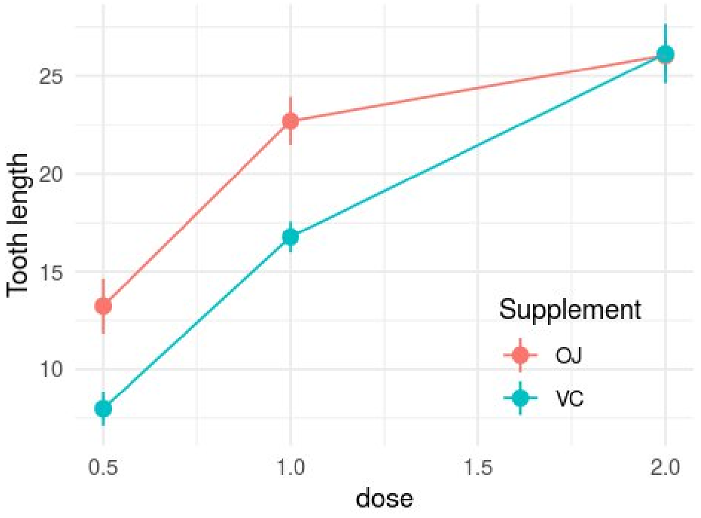
— a single variable (here, *dose*) can be encoded with **2** parameters

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*!

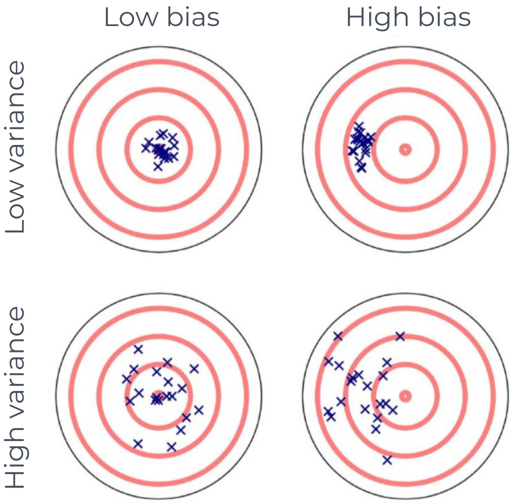


Bias-variance tradeoff

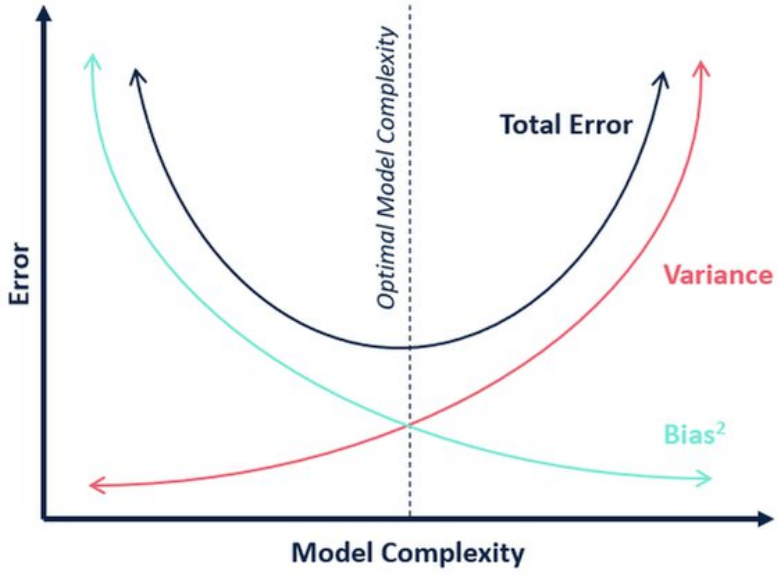
Bias = systematic error due to inadequate model (**underfitting**)

Variance = sensitivity to small fluctuations in the data (**overfitting**)

→ variability of parameter estimates across replications



Target center = true value
Crosses = model predictions



Bayes Factor

As an Occam's razor

$$BF_{21} = \frac{p\,(y|M_2)}{p\,(y|M_1)} = \frac{\int p\,(y|\theta_2)\,p(\theta_2)d\theta_2}{\int p\,(y|\theta_1)\,p(\theta_1)d\theta_1}$$

Model complexity is automatically **penalized** by the Bayes Factor: the more parameters, the more the prior is spread out over “irrelevant” regions, the more “diluted” the predictive power of the model

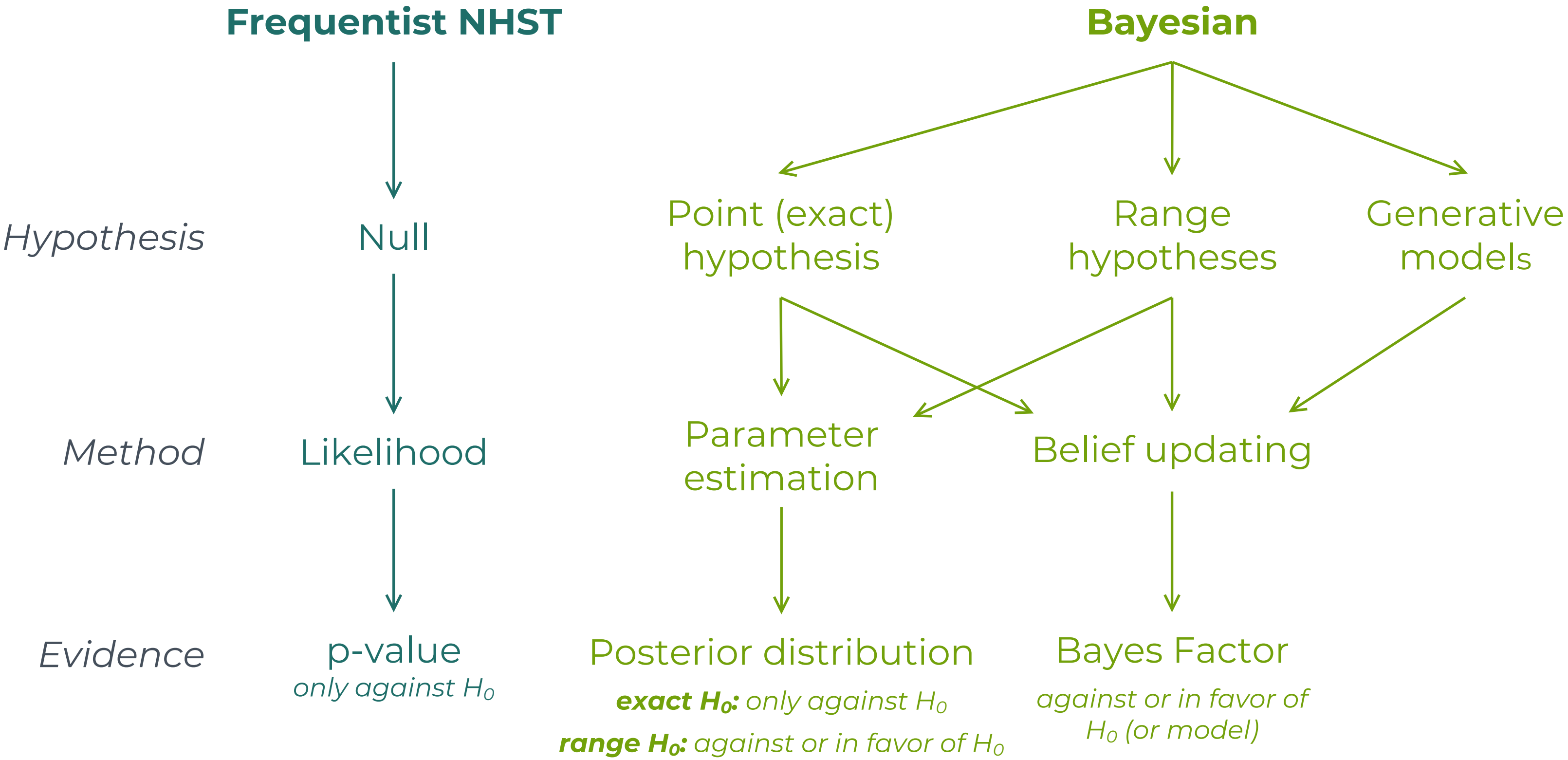
⇒ diffuse priors follow the same logic

Bayes Factor

Application to computational models

TO DO

The many ways of Bayesian hypothesis testing



Frequentist NHST

Bayesian

Hypothesis

Null

Point (exact) hypothesis

Range hypotheses

Generative models

Method

Likelihood

Parameter estimation

Belief updating

Evidence

p-value

only against H_0

Posterior distribution

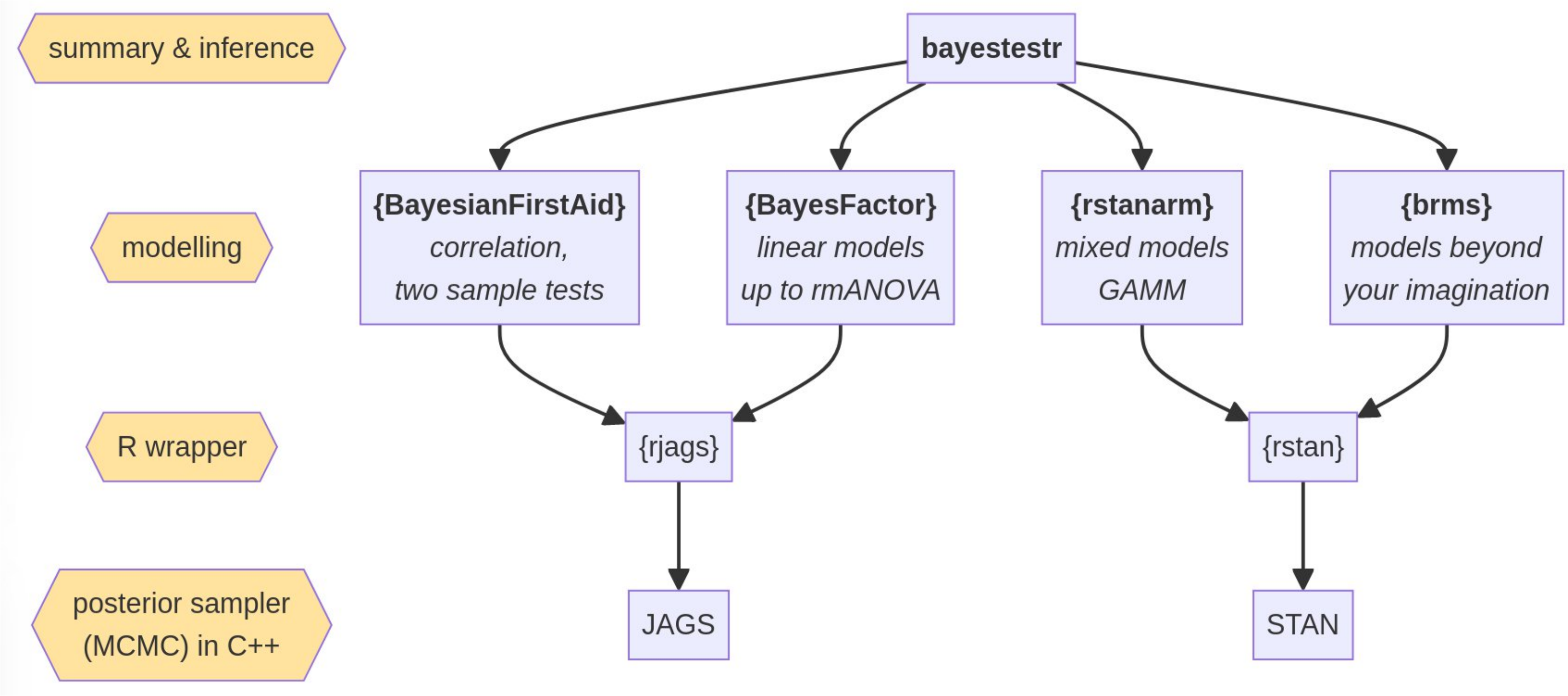
exact H_0: *only against H_0*

range H_0: *against or in favor of H_0*

Bayes Factor

against or in favor of H_0 (or model)

Software & package ecosystem


summary & inference
modelling
R wrapper
posterior sampler (MCMC) in C++
bayestestr
{BayesianFirstAid} correlation, two sample tests
{BayesFactor} linear models up to rmANOVA
{rstanarm} mixed models GAMM
{brms} models beyond your imagination
{rjags}
{rstan}
JAGS
STAN

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**