An introduction to Bayesian Data Analysis using Stan

Lionel Hertzog

Statistical workshop,Thünen Institut for Biodiversity, 27.11.2019

Structure of the talk

- What is Bayesian Data Analysis?
- How to do Bayesian Data Analysis?
- Why do Bayesian Data Analysis?

Bayesians can't walk on water

Except maybe shallow water

There is no bayesian model

What is Bayesian Data Analysis?

Posterior $\propto$ Likelihood * Prior

Or:

New knowledge $\propto$ new data * prior knowledge

Bayesian data analysis updates prior knowledge (or belief) based on new data.

The Likelihood: what we've been doing all of our lives

$$lm(y \sim x1 + x2, data)$$

This is equivalent to:

$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 * x1 + \beta_2 * x2$$

and

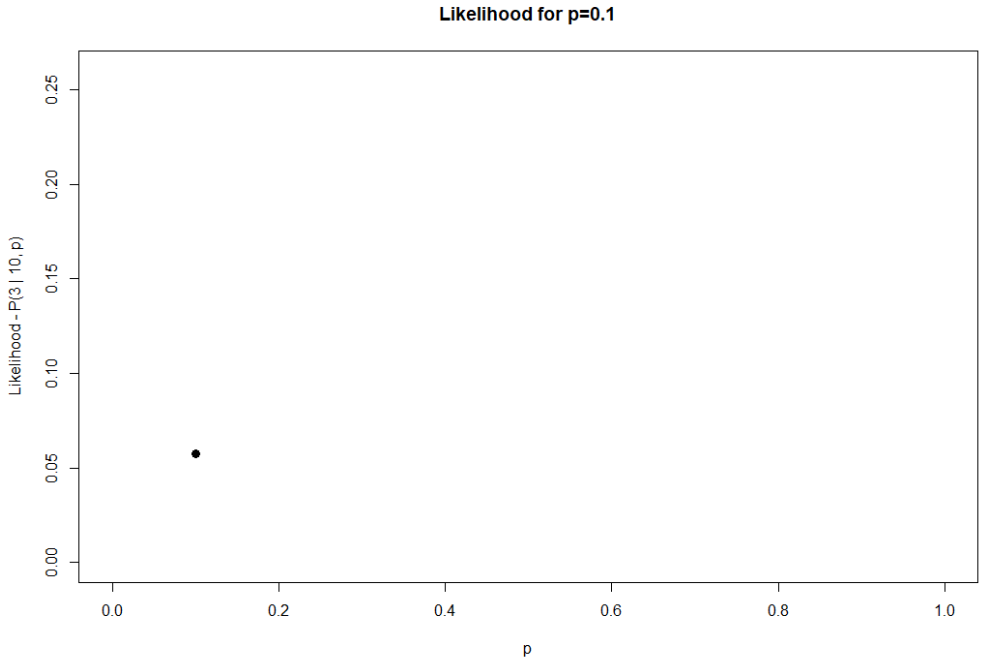
$$P(y|\beta_0, \beta_1, \beta_2, \sigma)$$

is the likelihood, the probability of the data given the model: Prob(data | model)

The likelihood: an example

How much earth is on earth?

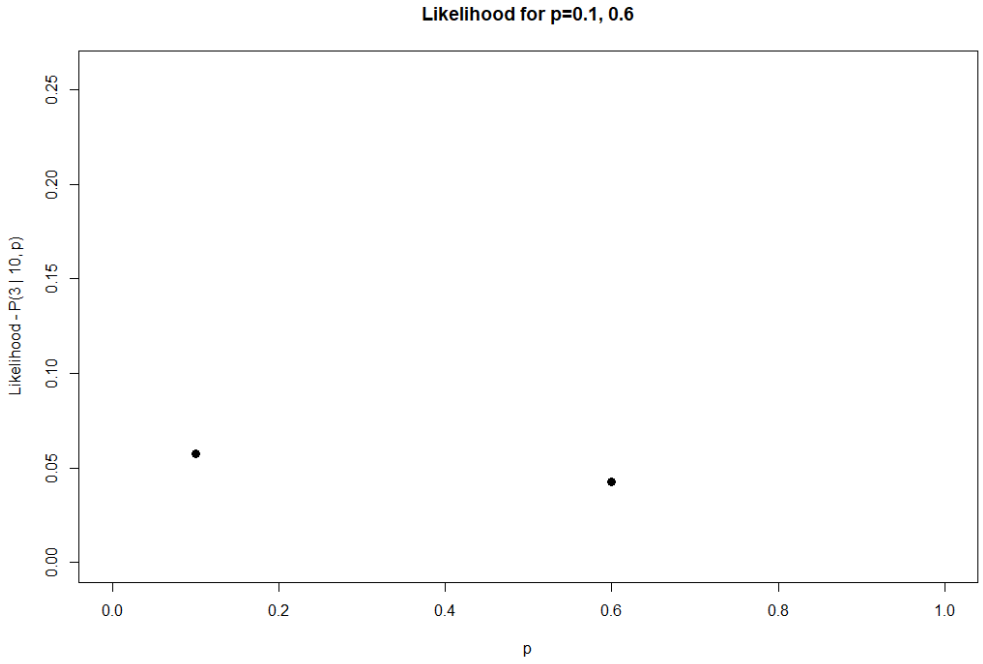
We tossed the globe 10 times and we landed 3 times on land. The likelihood is: $P(3 \mid 10, p)$, we use the binomial distribution



The likelihood: an example

How much earth is on earth?

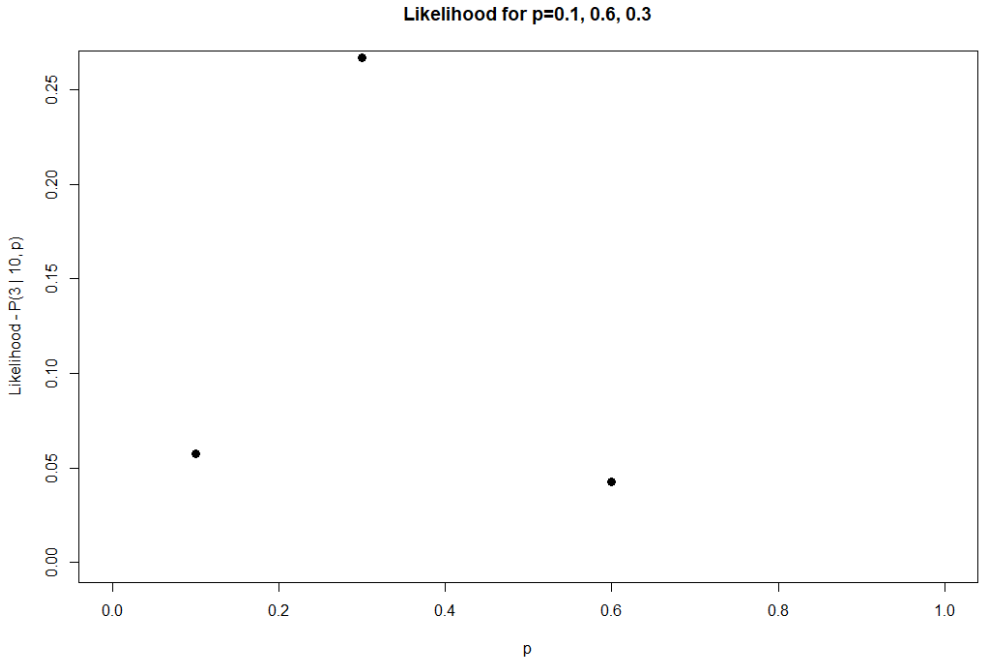
We tossed the globe 10 times and we landed 3 times on land. The likelihood is: $P(3 \mid 10, p)$, we use the binomial distribution



The likelihood: an example

How much earth is on earth?

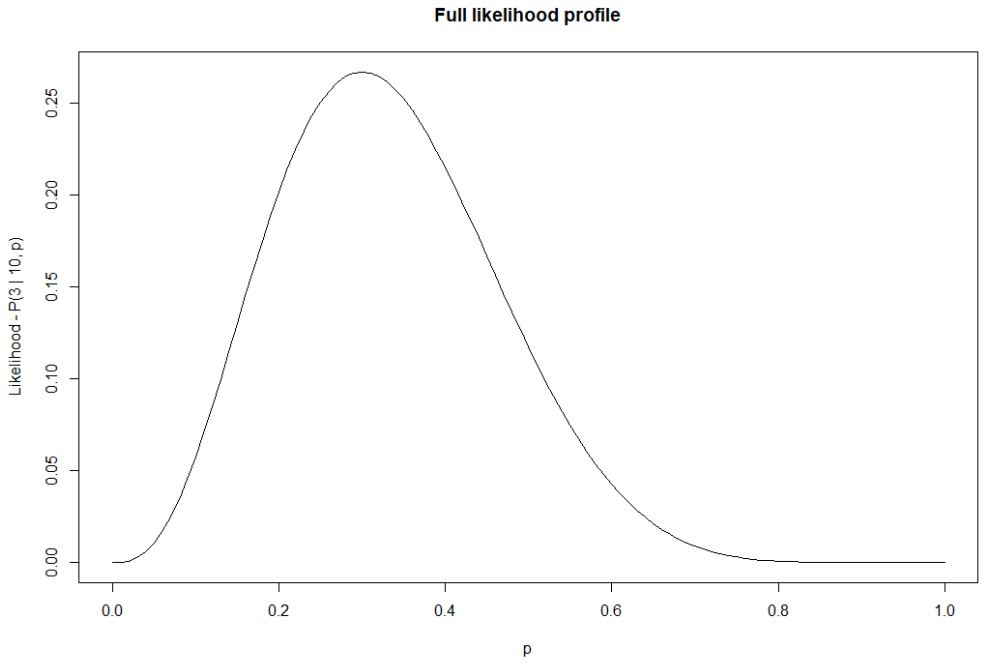
We tossed the globe 10 times and we landed 3 times on land. The likelihood is: $P(3 \mid 10, p)$, we use the binomial distribution



The likelihood: an example

How much earth is on earth?

We tossed the globe 10 times and we landed 3 times on land. The likelihood is: $P(3 \mid 10, p)$, we use the binomial distribution

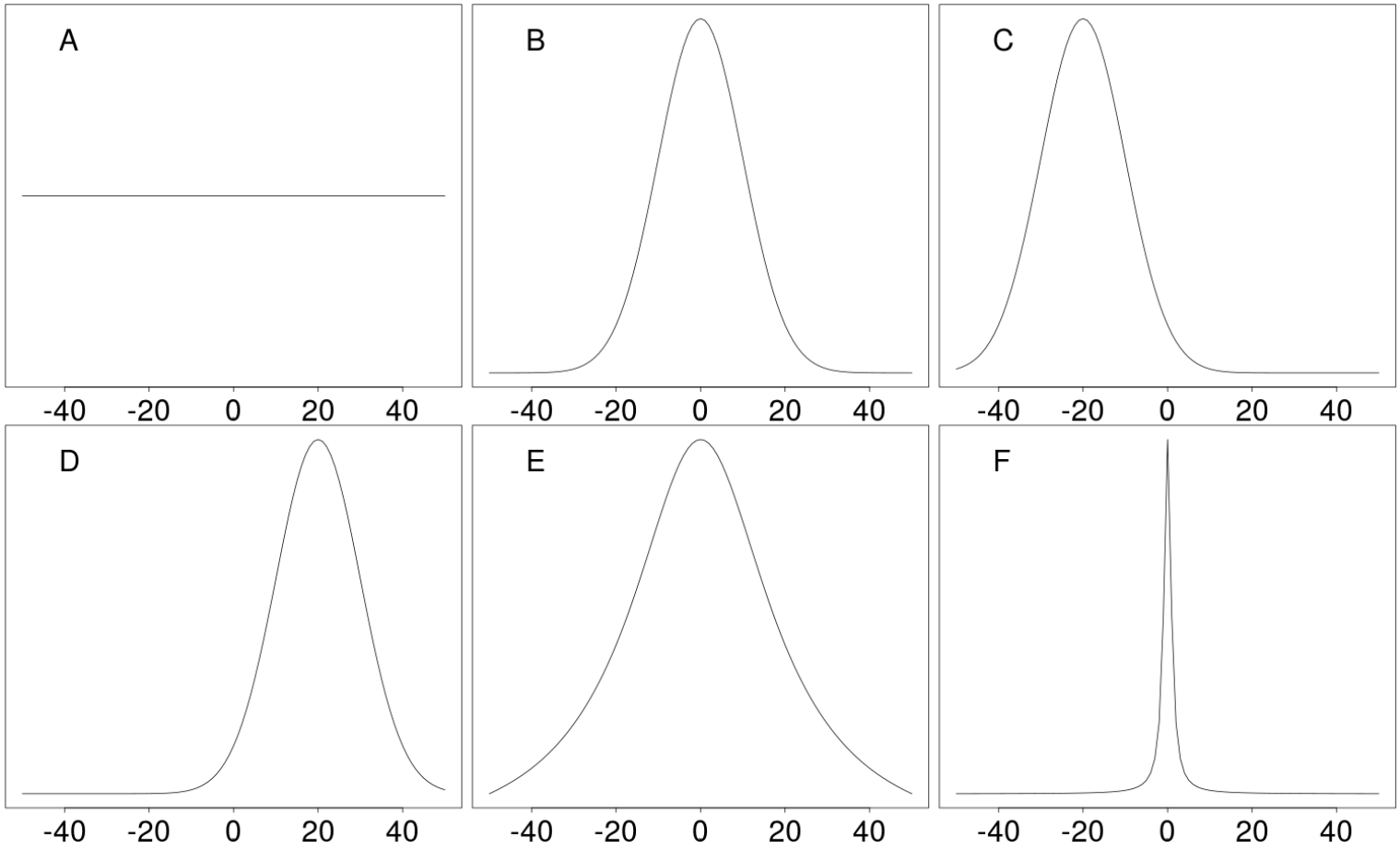


Exercice time : about likelihood

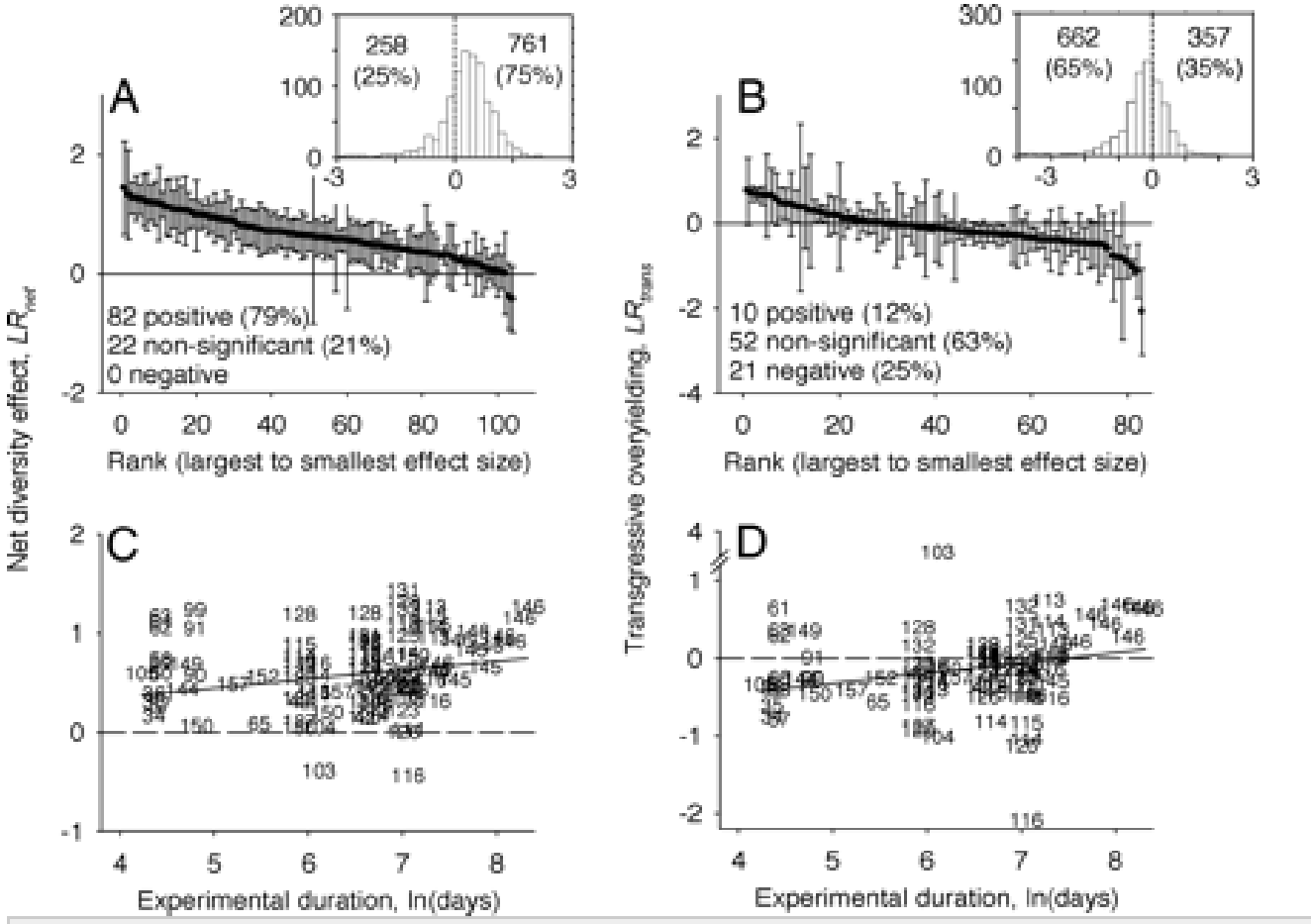
Do as many Exercice 1 as you can

The priors: our educated guesses about the world

Amongst the following probability distribution which one would you think represent plausible distribution of the slope of plant richness effect on plant biomass?



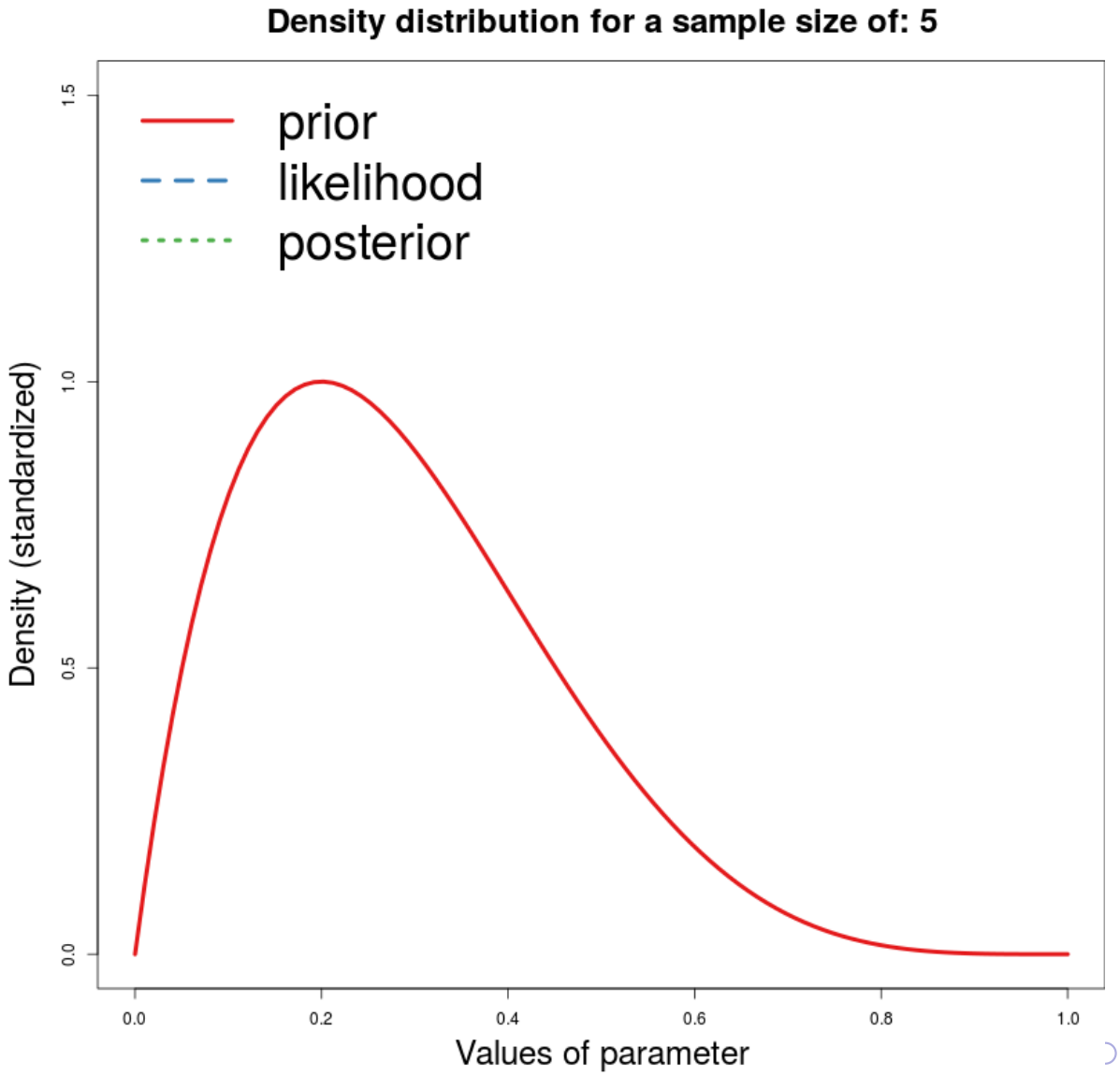
The priors: how to construct them



Cardinale et al, (2012) PNAS

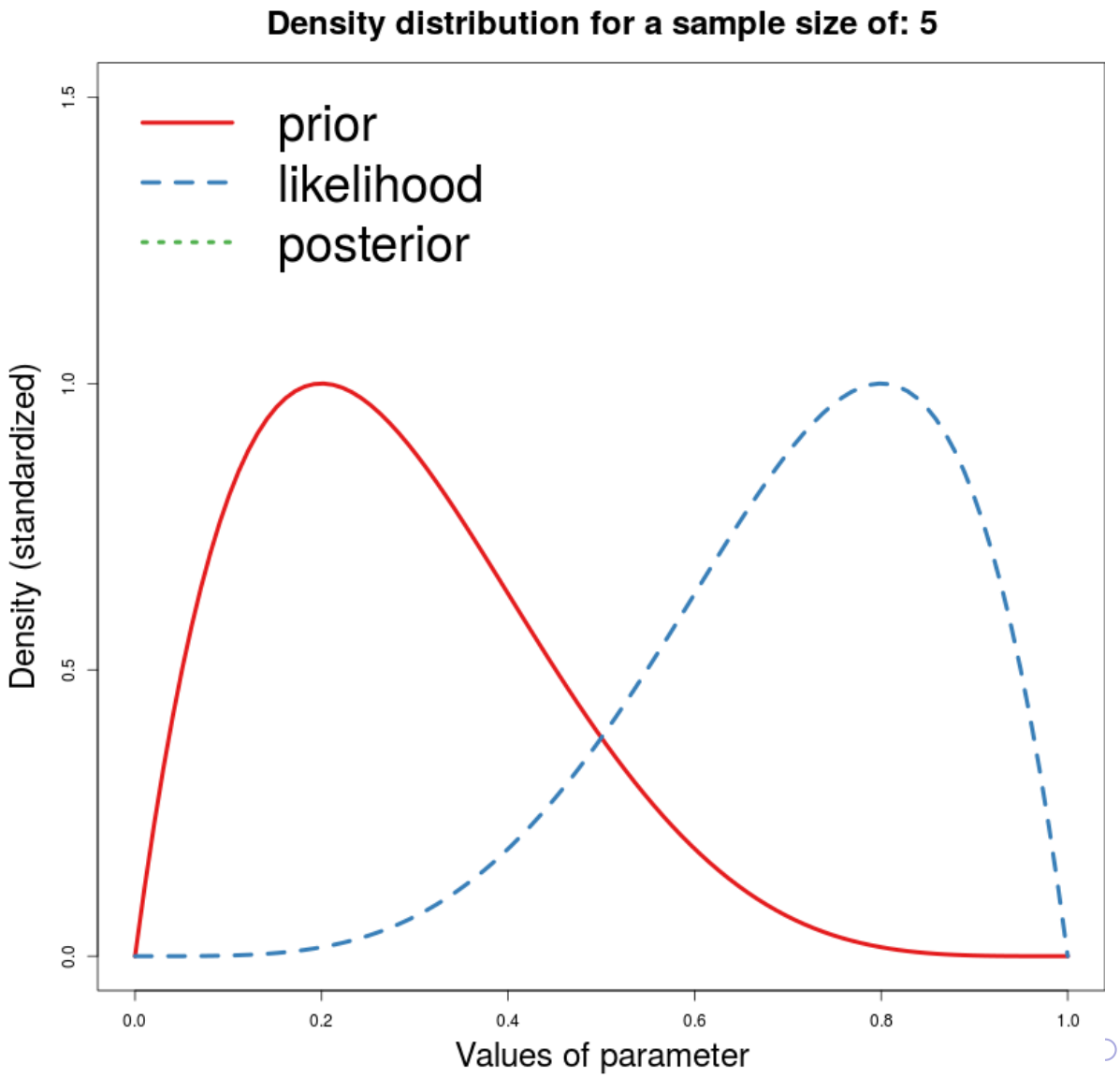
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- P(model | data)
- Weight of prior decline as sample size increases (for a given nb of parameters)



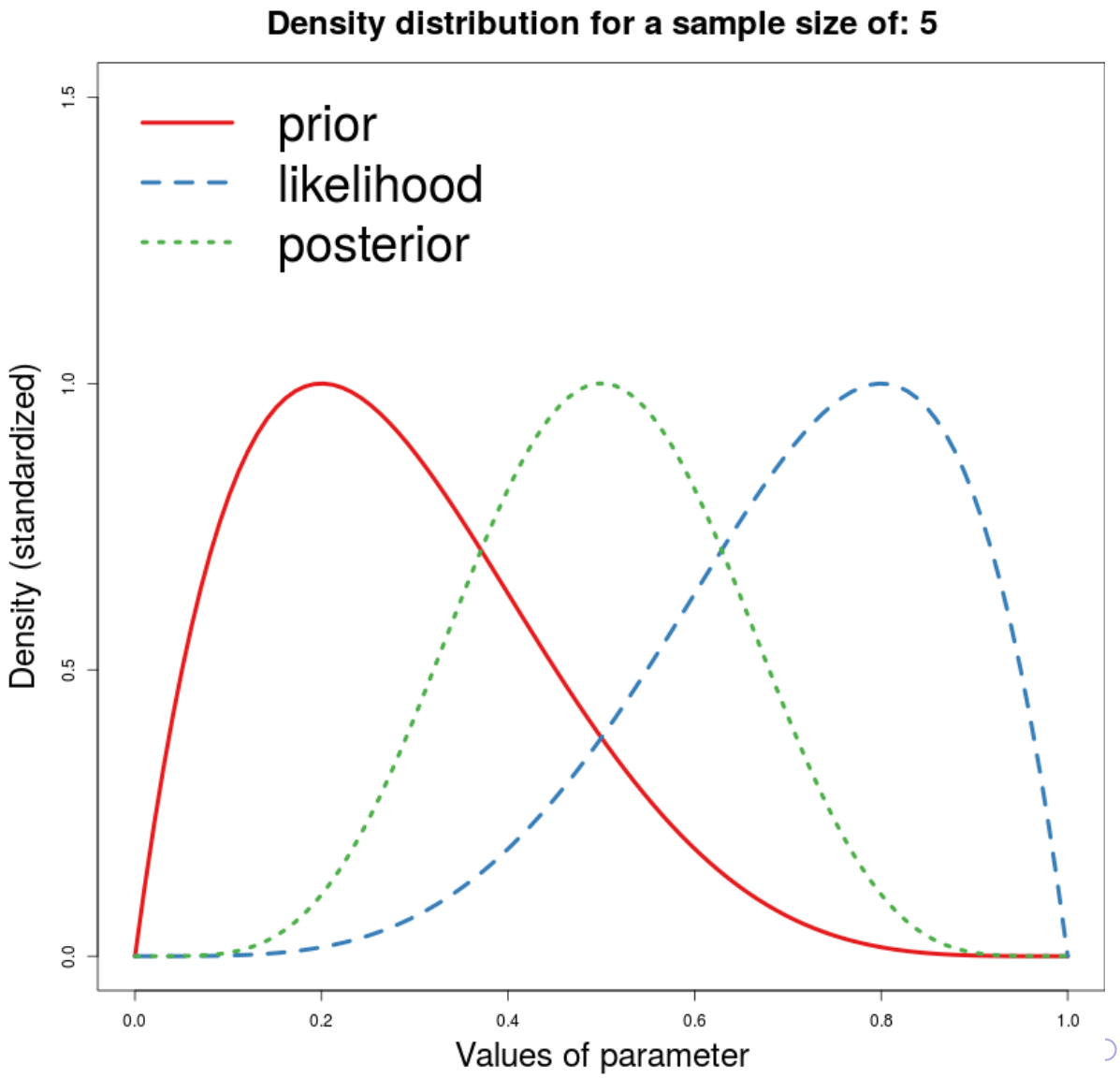
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- P(model | data)
- Weight of prior decline as sample size increases (for a given nb of parameters)



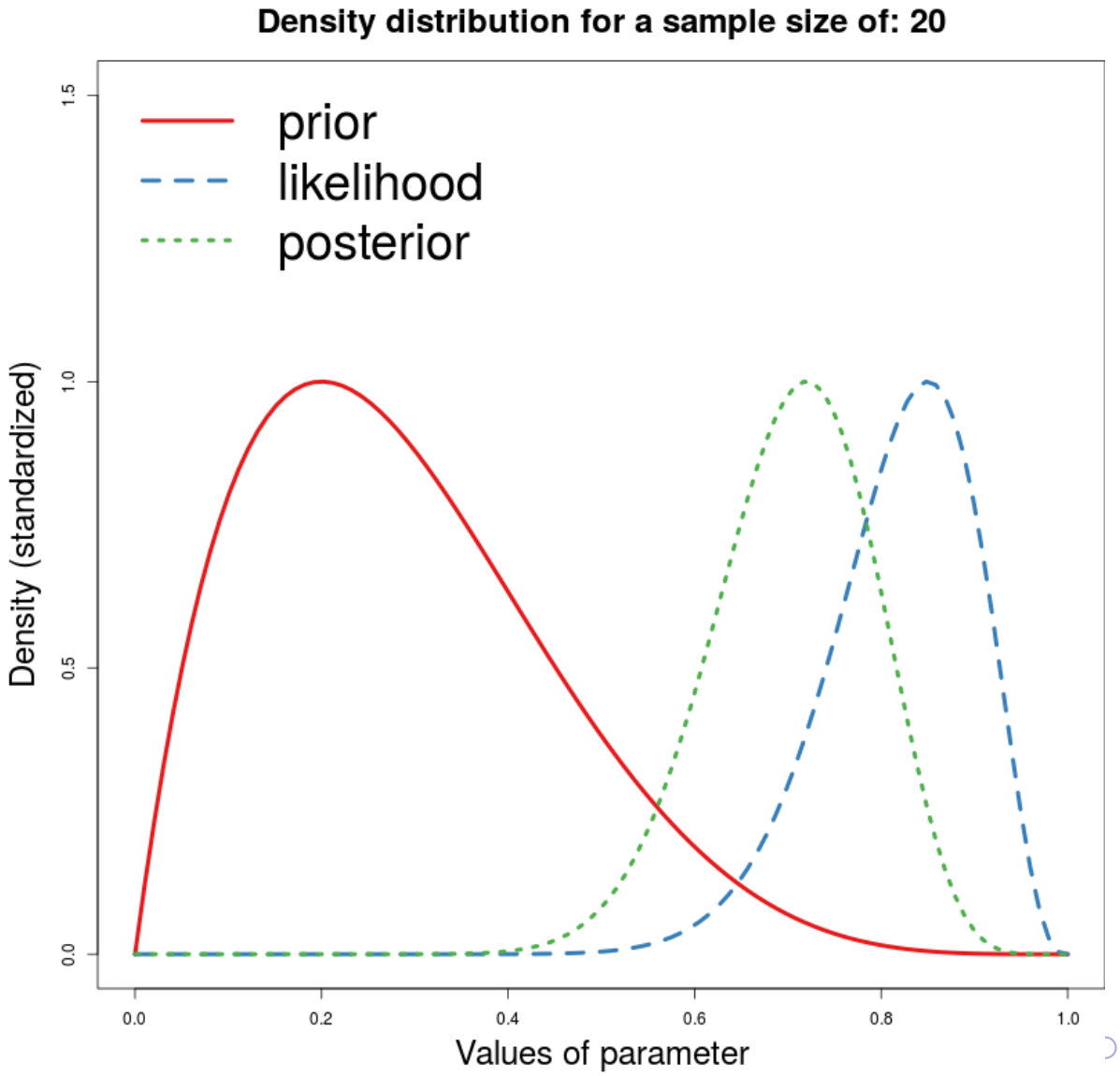
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- P(model | data)
- Weight of prior decline as sample size increases (for a given nb of parameters)



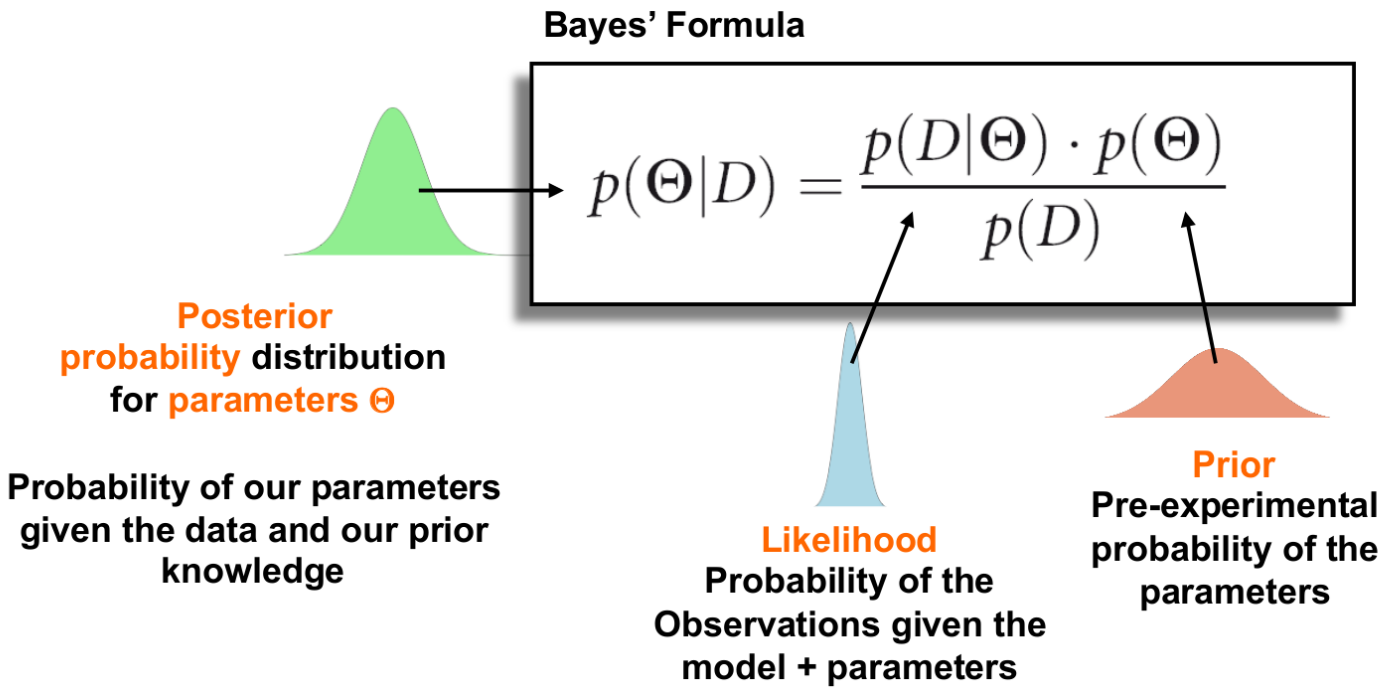
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- P(model | data)
- Weight of prior decline as sample size increases (for a given nb of parameters)



The key aspects of BDA

Hartig et al 2012 J. Veg. Sci.



- Everything is distribution
- Integrate prior knowledge
- Models as data-generators
- Easy interpretation

Ways to fit bayesian models in R

Two main options are available to fit bayesian models in R:

Directly using a dedicated probability language:

- JAGS via rjags
- Stan via rstan

Using packages that translate R formulas into the probability language:

- rstanarm
- brms
- INLA
- ...

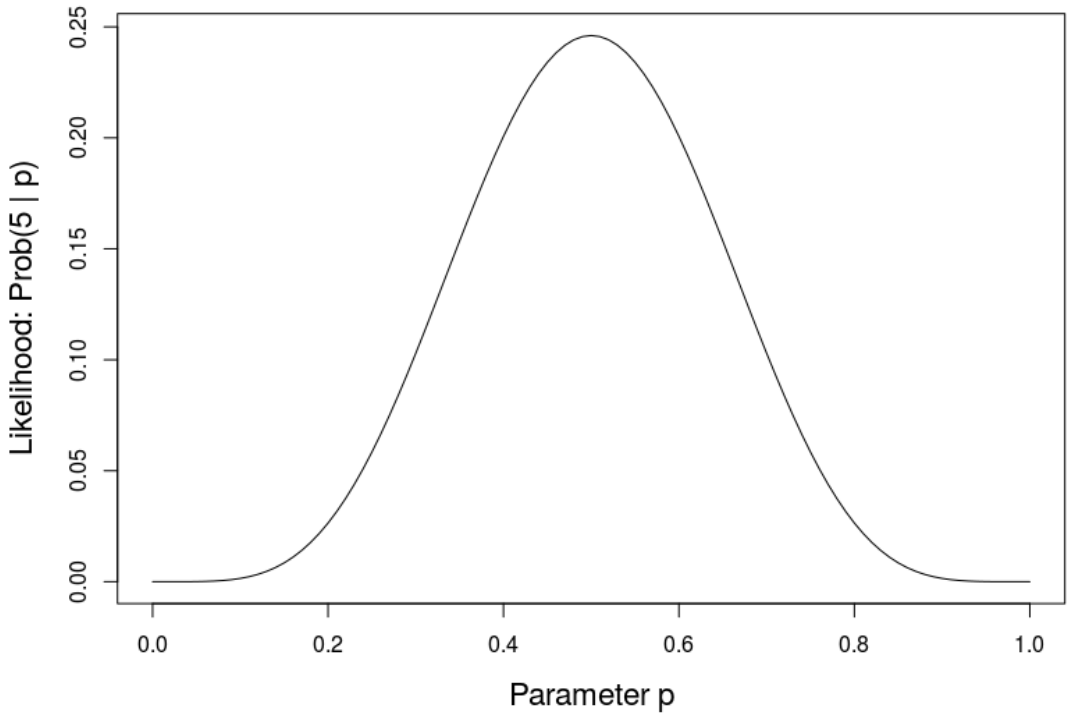
With these packages one can easily do Bayesian Data Analysis without needing to learn a new language.

Fit your first Bayesian model

Exercice time!

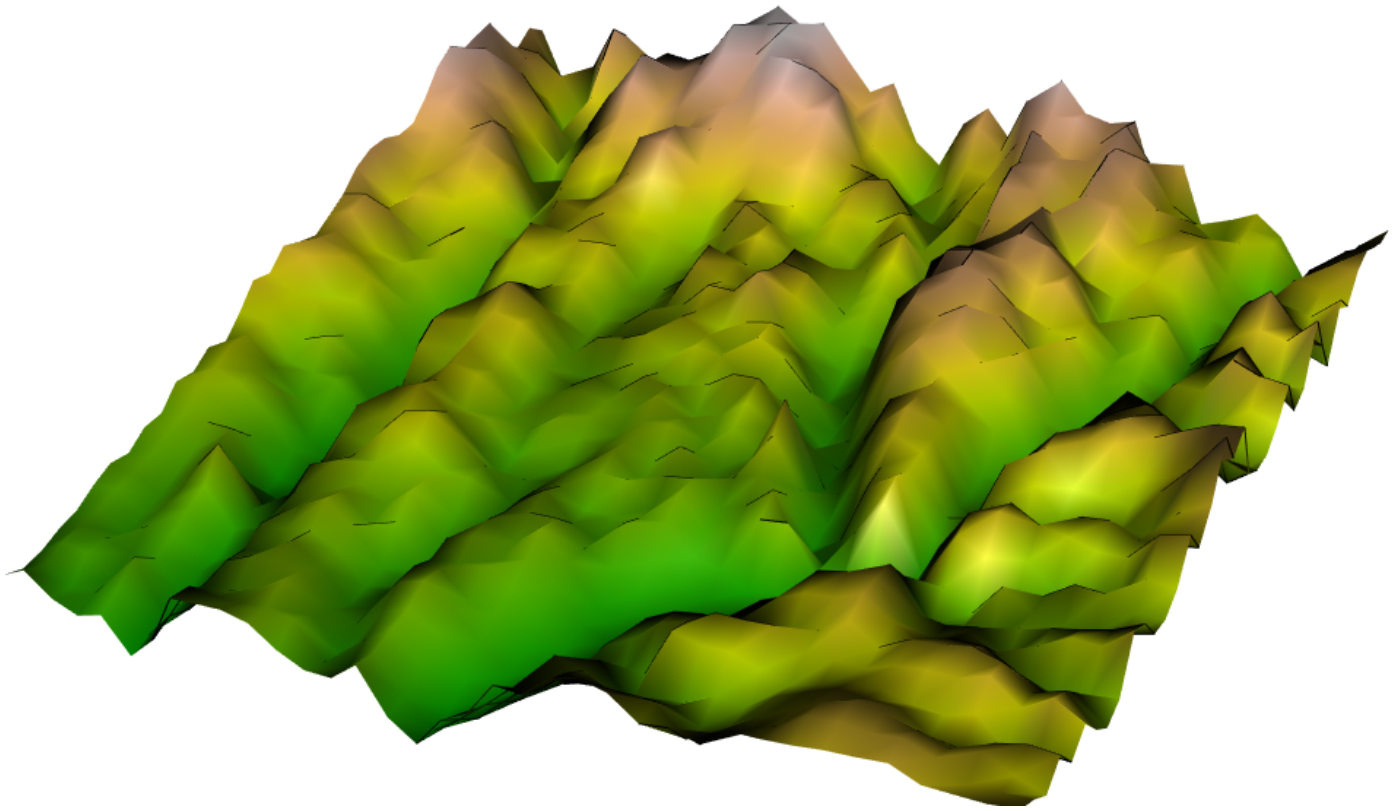
The sampling: the issue of complex likelihood surface

The likelihood surface if often (almost always) complex, how to get a reliable distribution?



The sampling: the issue of complex likelihood surface

The likelihood surface if often (almost always) complex, how to get a reliable distribution?

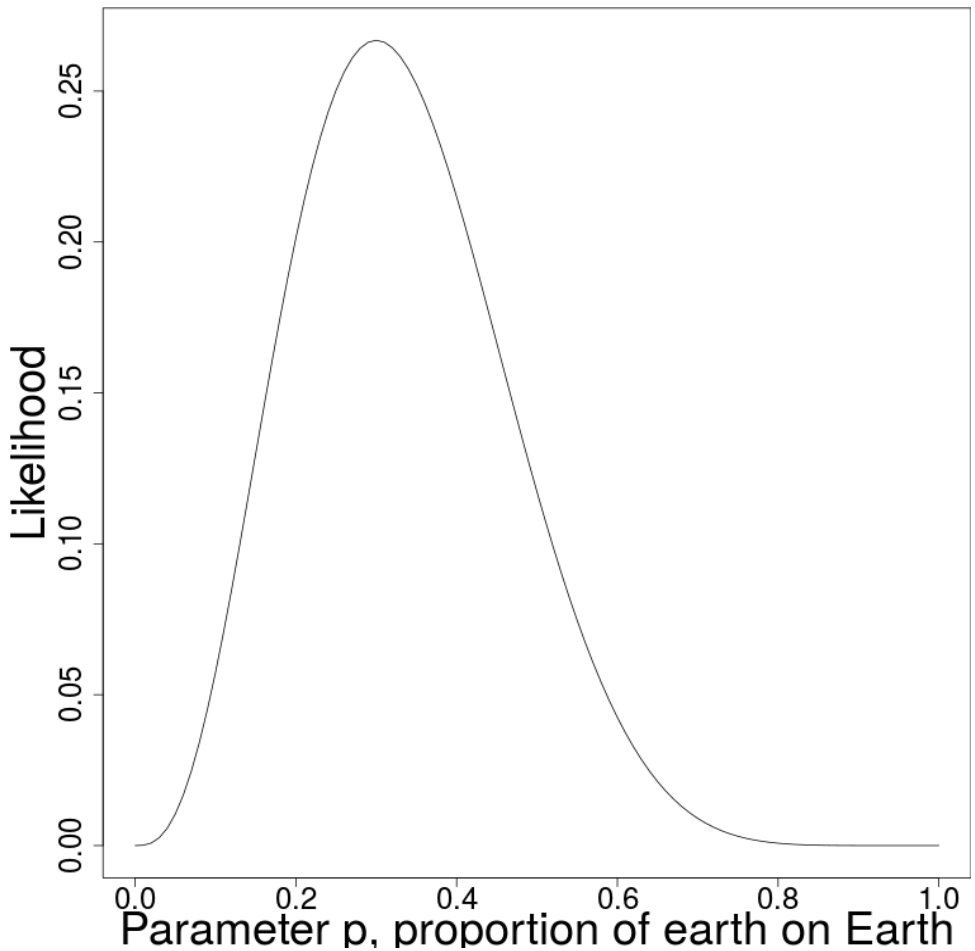


The sampler should travel around in the likelihood space but spend more time in area of high likelihood mass than in area of low likelihood mass

A simple MCMC example: How much earth is on Earth?

We tossed the globe 10 times and we landed 3 times on land. The parameter to estimate is: $\mathcal{B}(10, p)$, we assume flat prior.

(Example from Statistical rethinking, Richard McElreath)

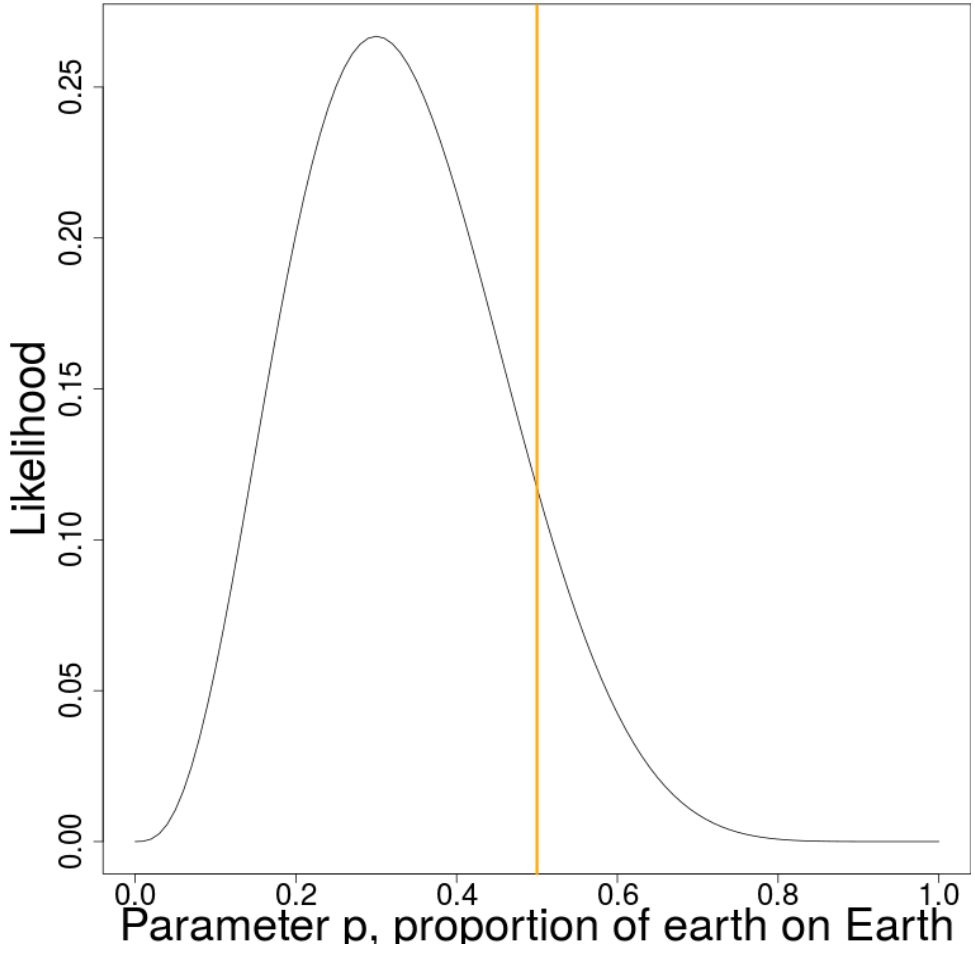


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5

Pick a starting value: 0.5, the likelihood is: 0.12

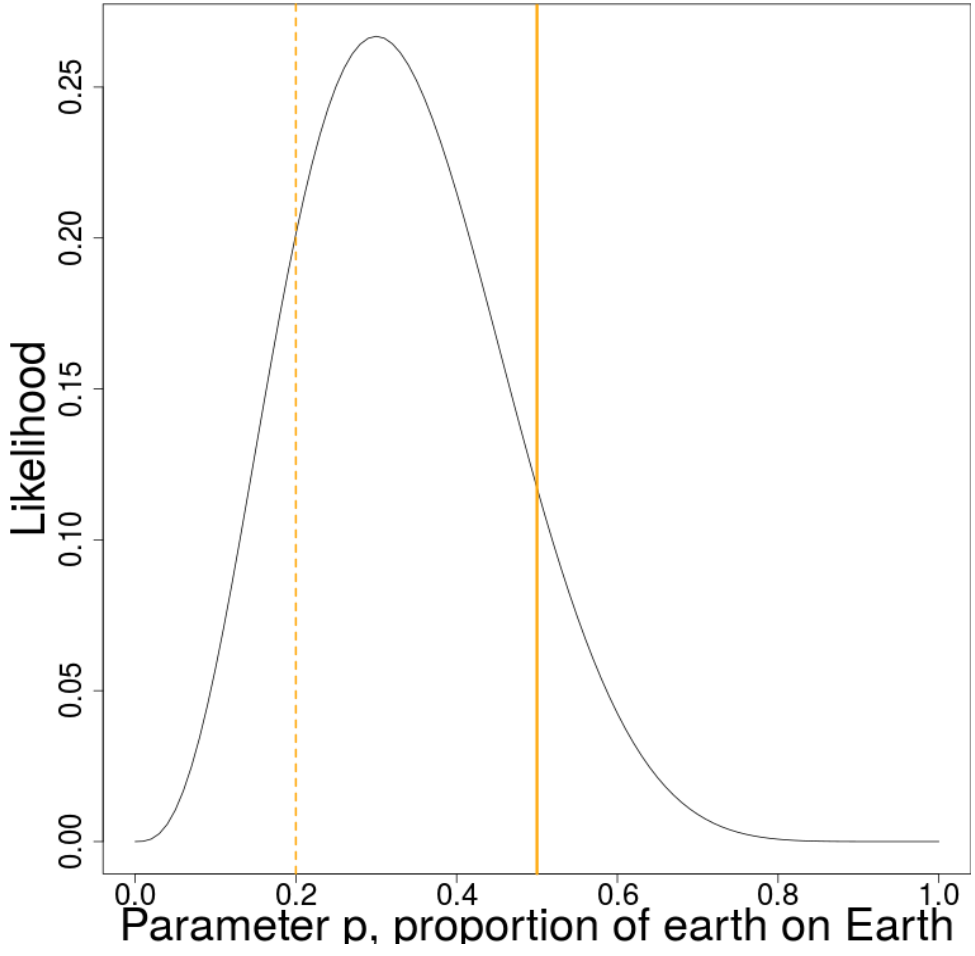


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5

Pick a new value: 0.2, the new likelihood is: 0.20

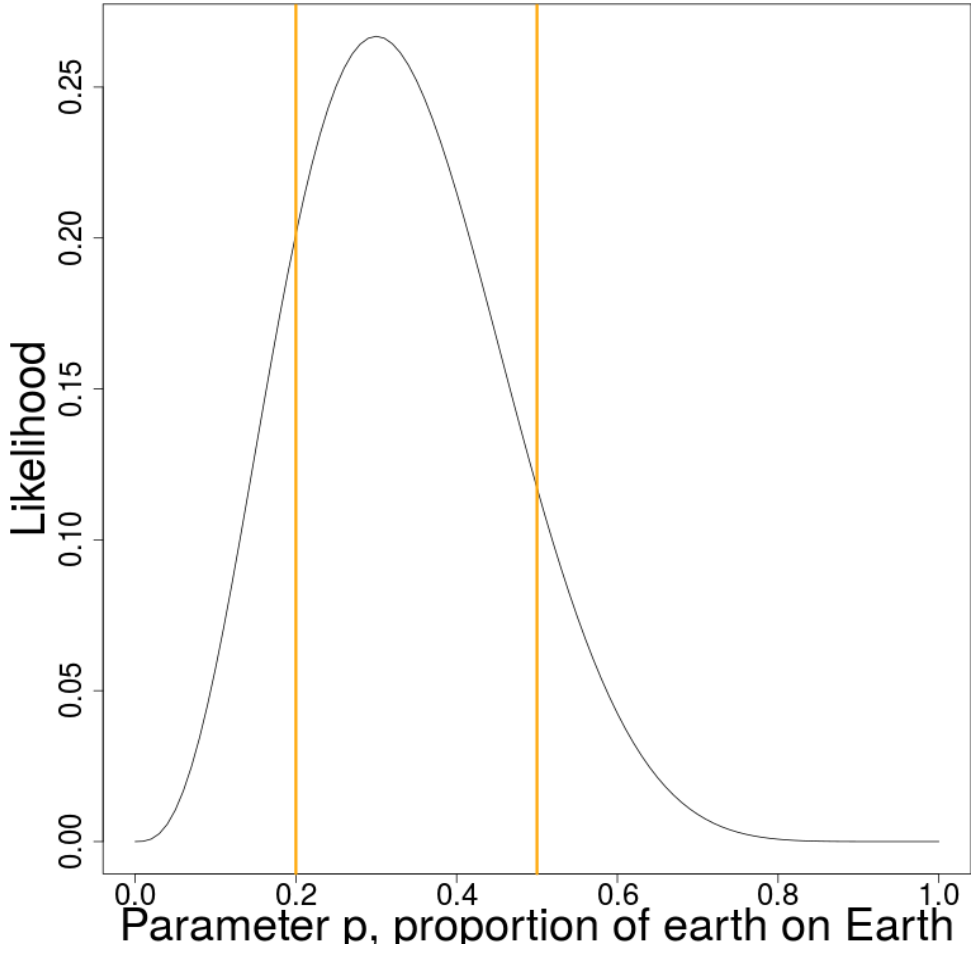


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5, 0.2

Old likelihood < New likelihood, jump.

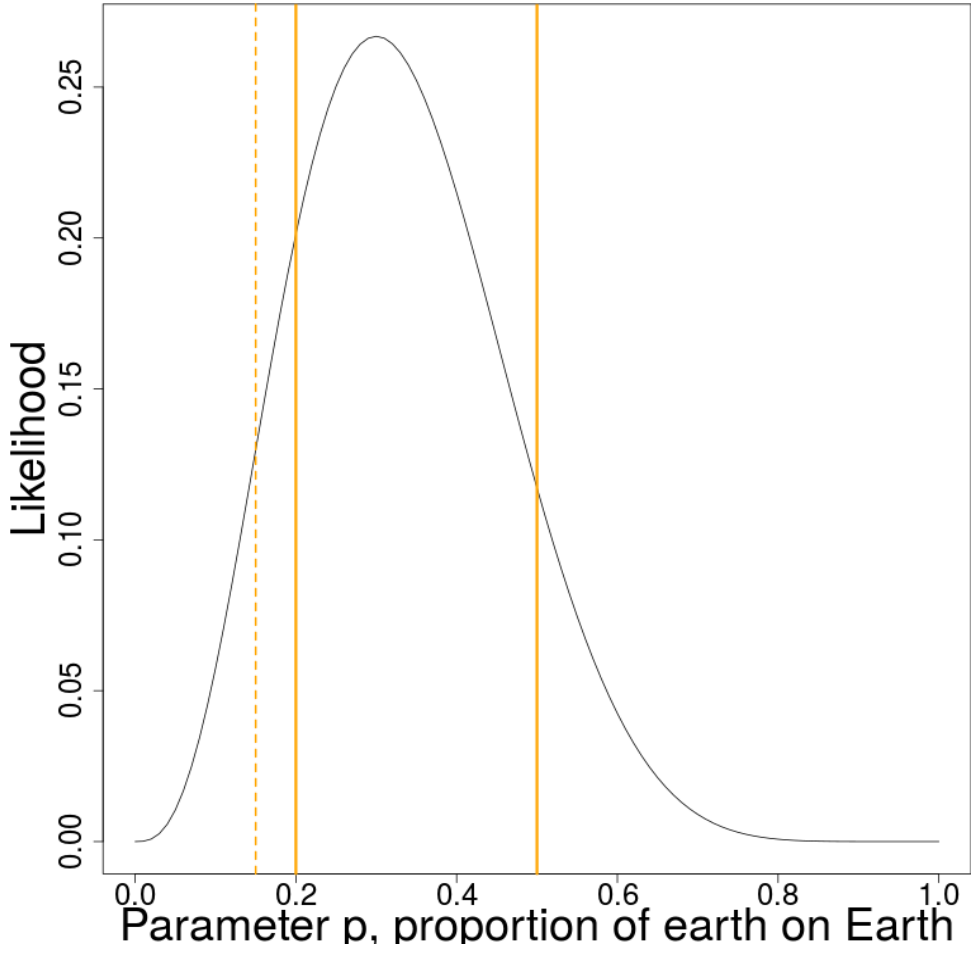


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5, 0.2

Pick a new value: 0.15, the new likelihood is: 0.13

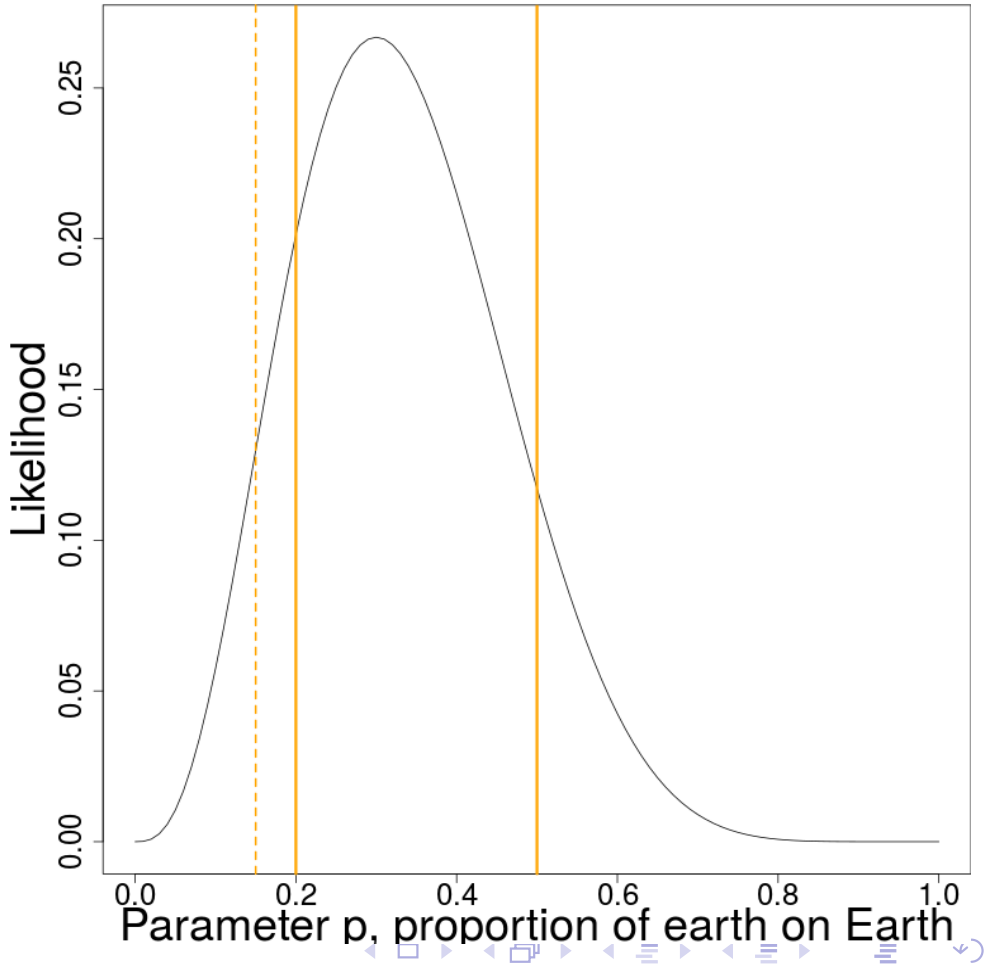


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5, 0.2

The new value will be accepted with a probability of 0.13 / 0.20

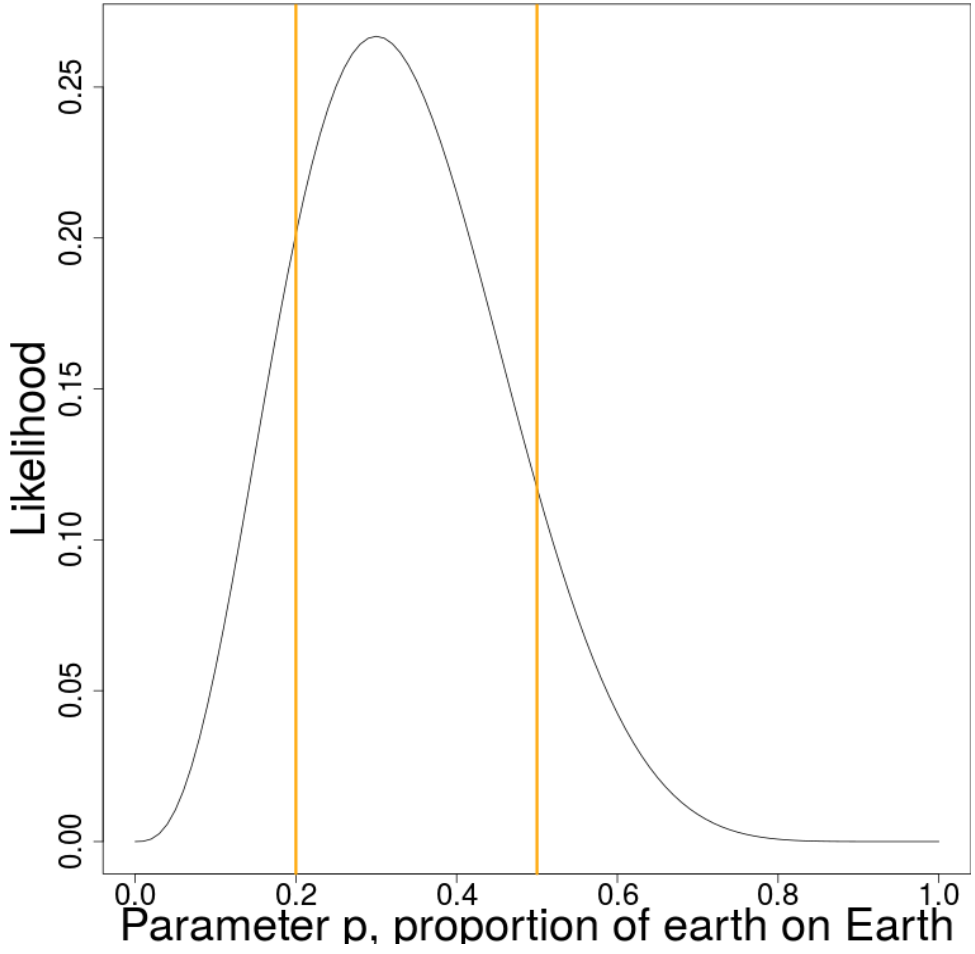


A simple MCMC example: How much earth is on Earth?

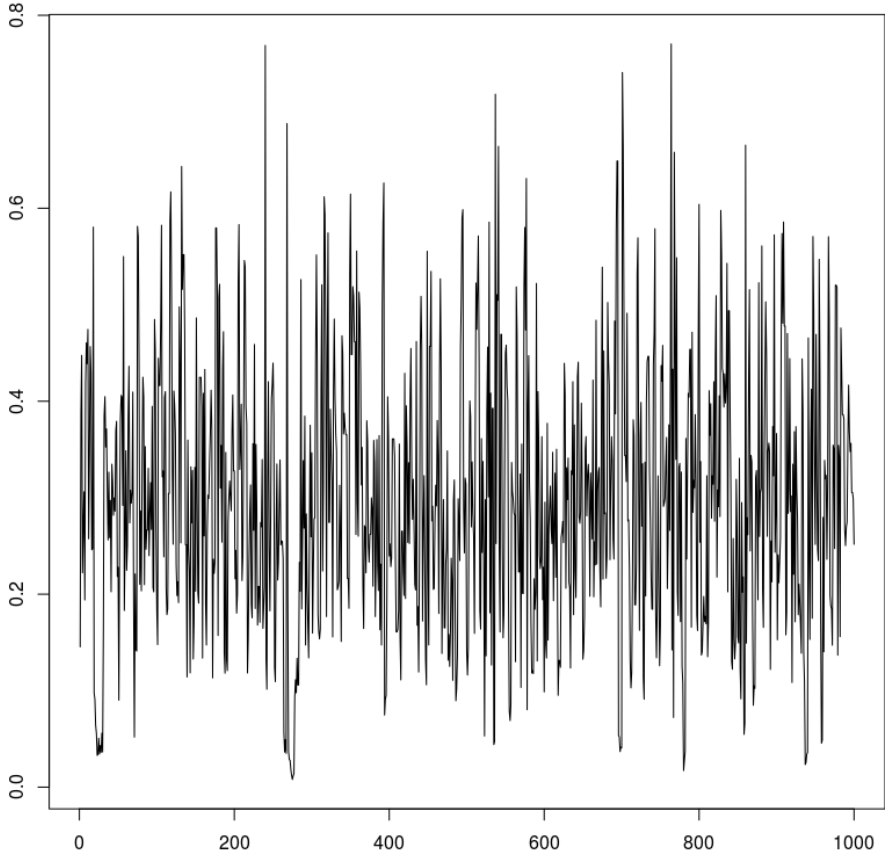
MCMC samples:

0.5, 0.2

The jump failed, go back to previous value



A simple MCMC example: How much earth is on Earth?

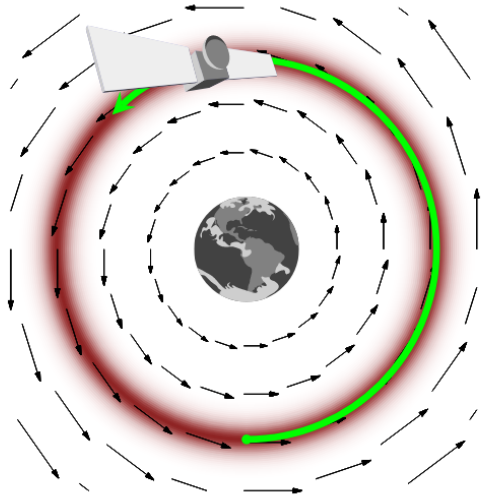


MCMC samples:

0.5, 0.2

MCMC sampling in real life

JAGS and Stan use different samplers, both have strength and weaknesses.



Stan uses Hamiltonian Monte-Carlo sampling which can be thought as sending a satellite with some momentum to sample around earth. Michael Betancourt (2017) `https://arxiv.org/abs/1701.02434`

More informations on sampling in JAGS and Stan:

- `https://www.youtube.com/watch?v=VnNdhsmOrJQ`
- `http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12681/full`

Elements of Bayesian vocabulary

- Chains: number of Markov chains ran, best to have at least 3
- Convergence: property of the Markov chains, at convergence the MCMC samples represent the posterior distribution
- Divergence: property of the Markov chains, when the sampler does not effectively move in the parameter space, in Stan it specifically means that the Hamiltonian dynamics ran into that indicates potential bias in estimates.
- Rhat: indicator to check convergence of the Markov chains, a value of 1 indicate convergence
- n_eff: number of effective samples, due to autocorrelation in the Markov chains fewer samples are taken than expected, n_eff / number of MCMC samples should be larger than 0.1

Model checking in Bayesian data analysis

Exercice time!

Model comparison/selection

A couple of information criteria metrics should be used:

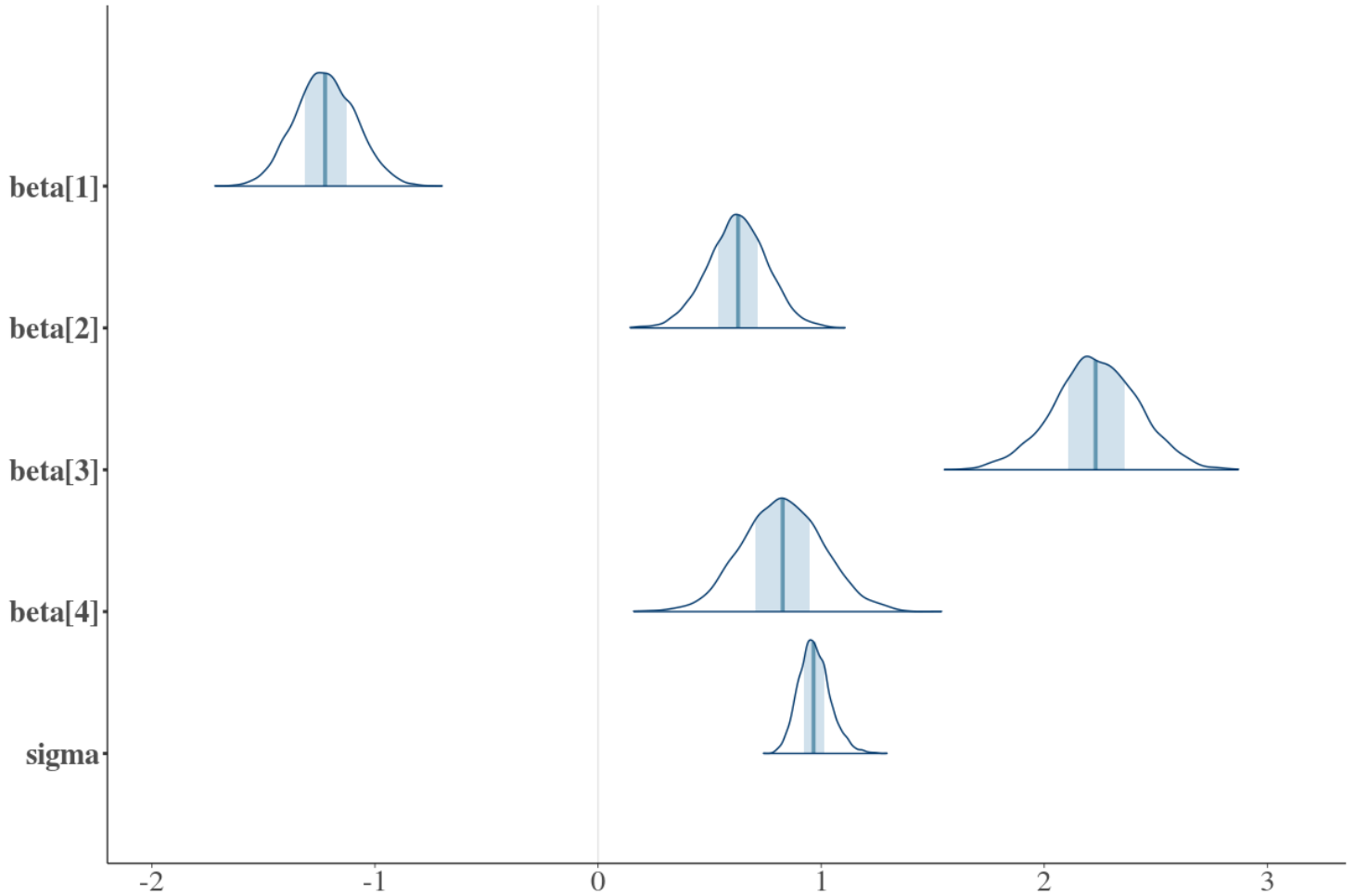
1. Watanabe-Akaike Information Criteria: basically the summed log likelihood over the posterior samples (predictive density) minus the effective number of parameters, better than the DIC since it uses all posterior samples instead of point estimates.
2. Leave-One Out cross-validation: drop one data point at a time and re-estimate the predictive density, this methods is commonly used for machine learning models to avoid problems like overfitting.

Both are readily available for Stan models through the **loo** package

Difference between frequentist and bayesian inference

- Frequentist approaches are based on the **likelihood** (probability of the data given the parameters) only, inference statement are based on imaginary repetition of the data collection.
- Bayesian approach is based on the **posterior** (probability of the model given the data and the prior), inference statement can be interpreted in terms of probability of the parameters.

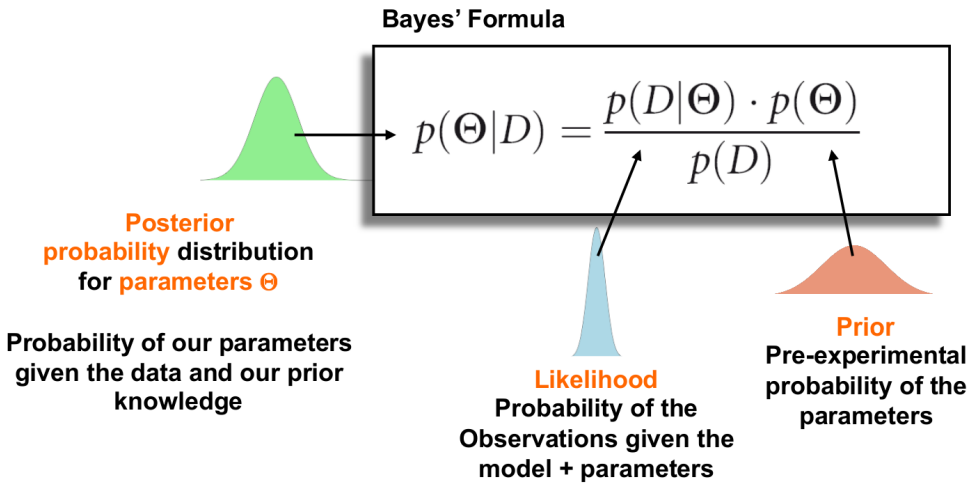
Embracing uncertainty



Flexibility in model building

It all comes down to the likelihood, as long as you can write down the likelihood function you can fit whatever model you want.

Hartig et al 2012 J. Veg. Sci.



Models as data storytelling

Once upon a time
in a lab / field site far, far away ...

- BDA allow great freedom in model building taking into account, if needed, sampling design, measurement processes ...
- Rather than squeeze your data to fit specific model assumptions, in BDA the models are adapted to the data

Bayesian approach output is what we actually want

Posterior samples from the MCMC can be interpreted as probabilities.

It is easy and straightforward to manipulate them to get what you want (probability intervals, hypothesis tests ...) all with easy interpretation.

Very different from complex and convoluted concepts like p-values, confidence intervals, null hypothesis ...

Asymptotic convergence, when bayesian and frequentist approach give similar answers

As sample size increase the posterior is drawn closer and closer to the likelihood, in other words at infinite sample size the posterior is the likelihood

The relative importance of the likelihood vs the prior depends on the complexity of the models, if sample size is fixed, the more parameters, the more the posterior is affected by the prior

Time for some discussion

Bayesian vs frequentist approach

Model to explore in next session?