An introduction to Bayesian Data Analysis using STAN

Lionel Hertzog & Maxime Dahirel

BDA workshop, Ecology Across Border meeting
13th December 2017, Gent

Structure of the workshop

1. General introduction to Bayesian Data Analysis (circa. 30min)
2. Examples of BDA workflow (circa. 30min)
3. Small group discussion on specific themes (circa. 30min)

Make sure to check the github page of the workshop:
`https://github.com/lionel68/STAN-BES-2017`

Structure of the talk

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

Why do we do Stats?

1. To test hypothesis?
2. To explore patterns and relationships?
3. To predict / forecast to new place and time?

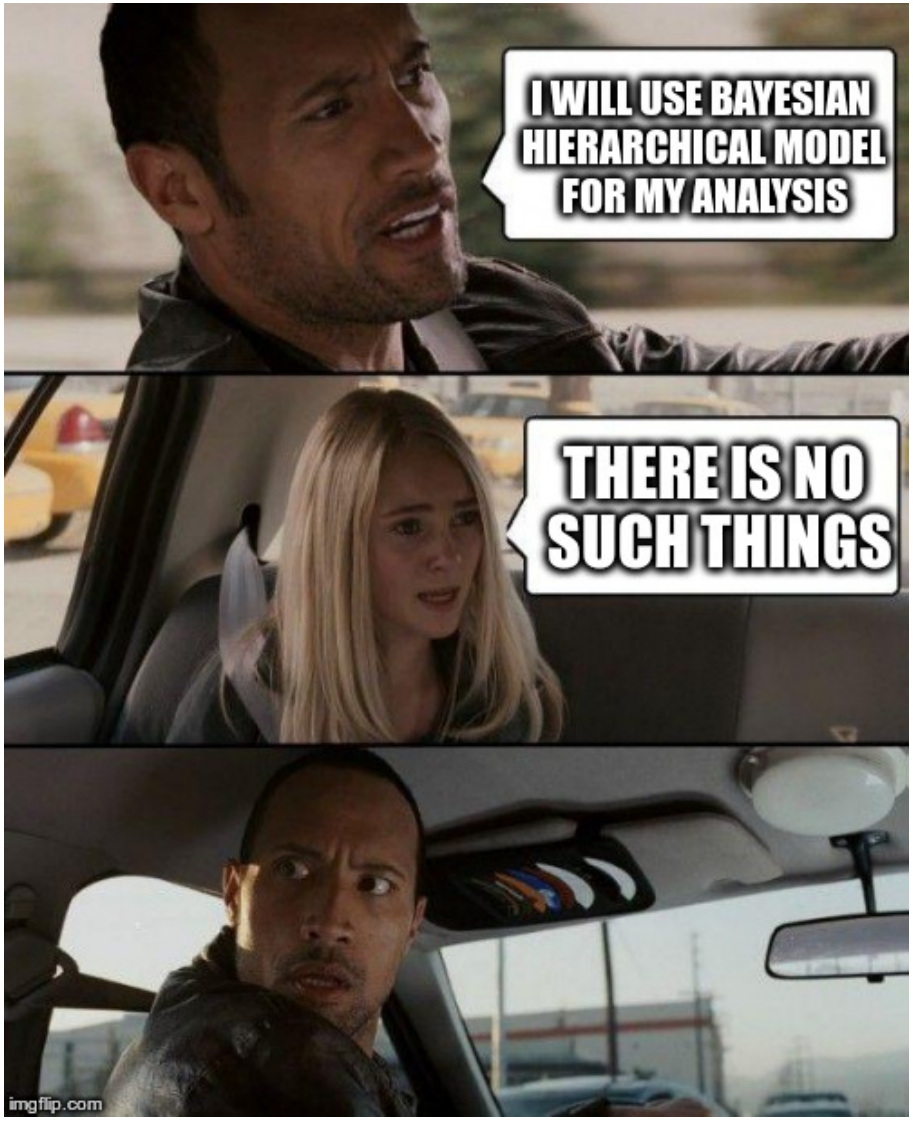
More reasons to do stats?

Bayesian data analysis is one way to do stats, it is well suited for some situation but poorly suited for others.

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

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

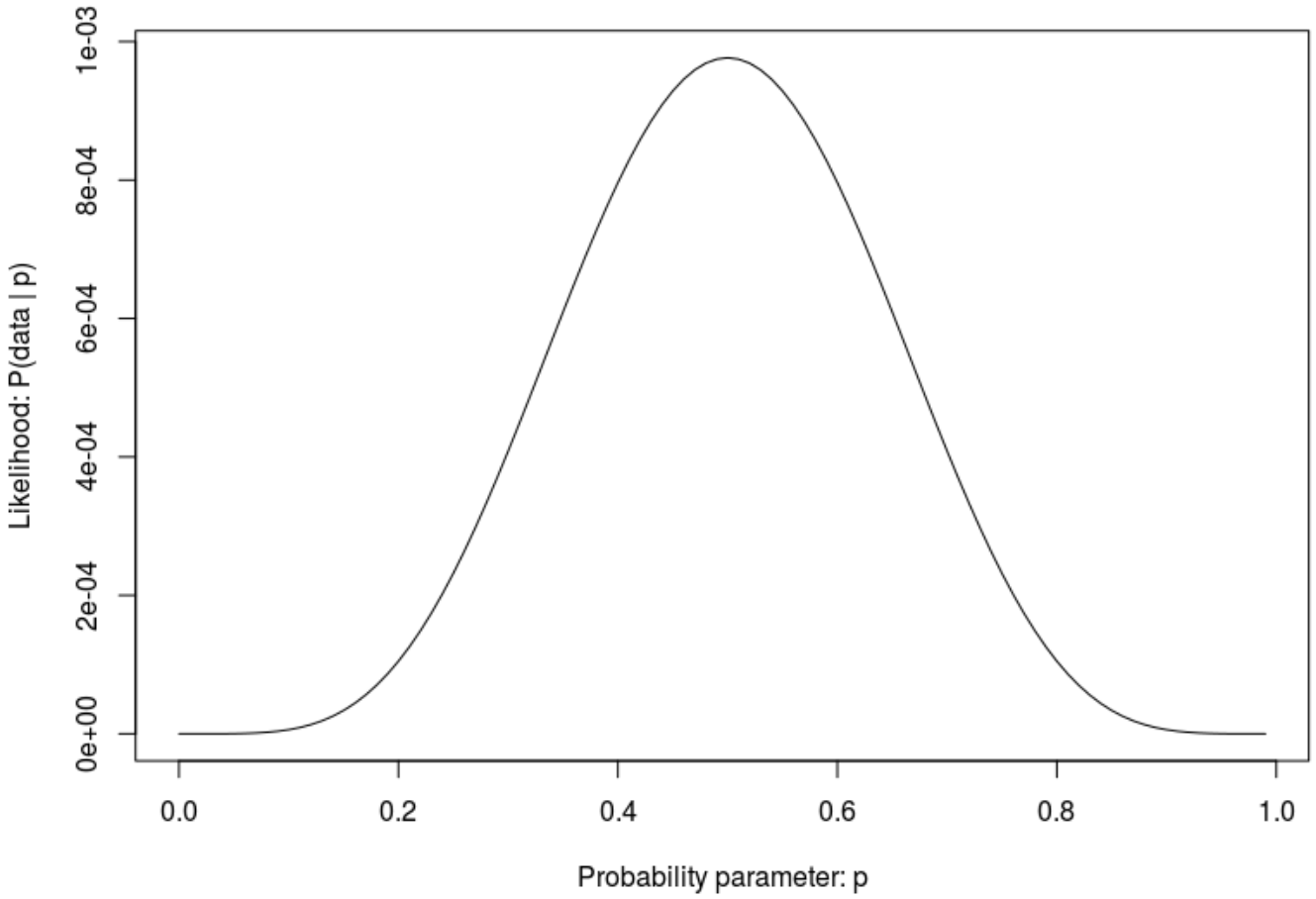
This is equivalent to:

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

$$\mu_i = a + b_1 * x1 + b_2 * x2$$

Which is called the LIKELIHOOD, the probability of the data given the model.

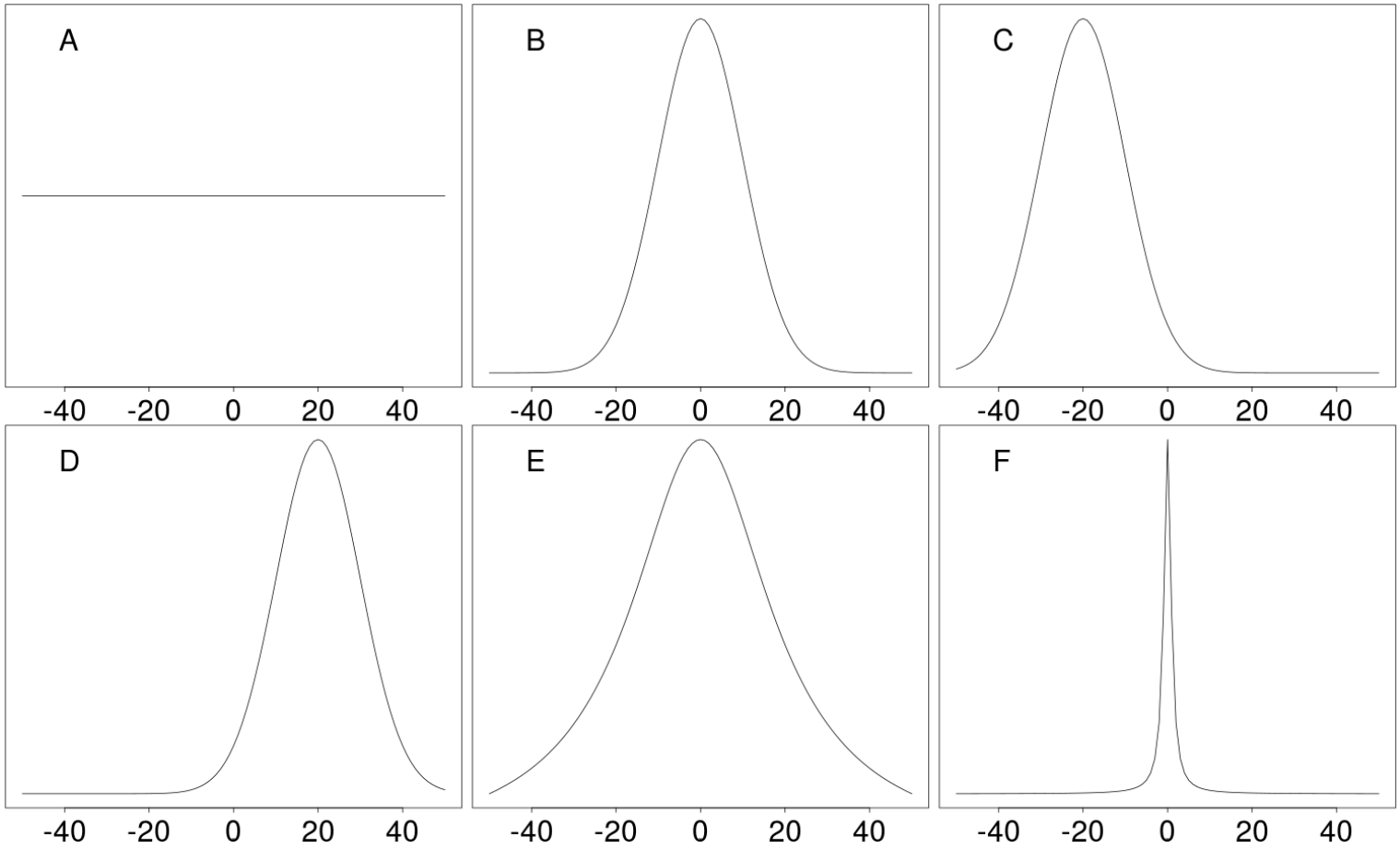
The likelihood: a graphical example



$$y \sim \mathcal{B}(n = 10, size = 1, prob = 0.5)$$

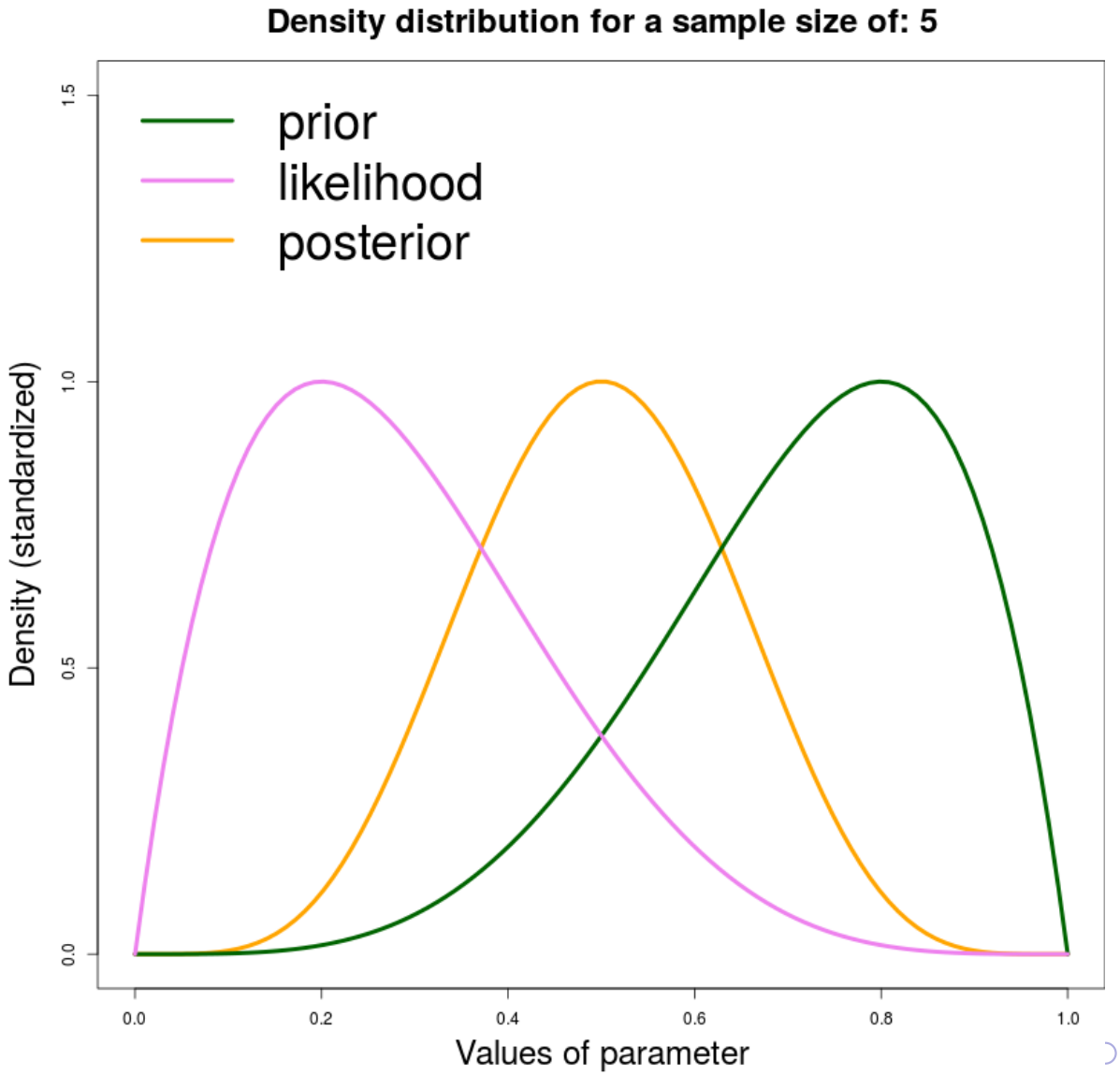
The prior: our eductaed guess about the world

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



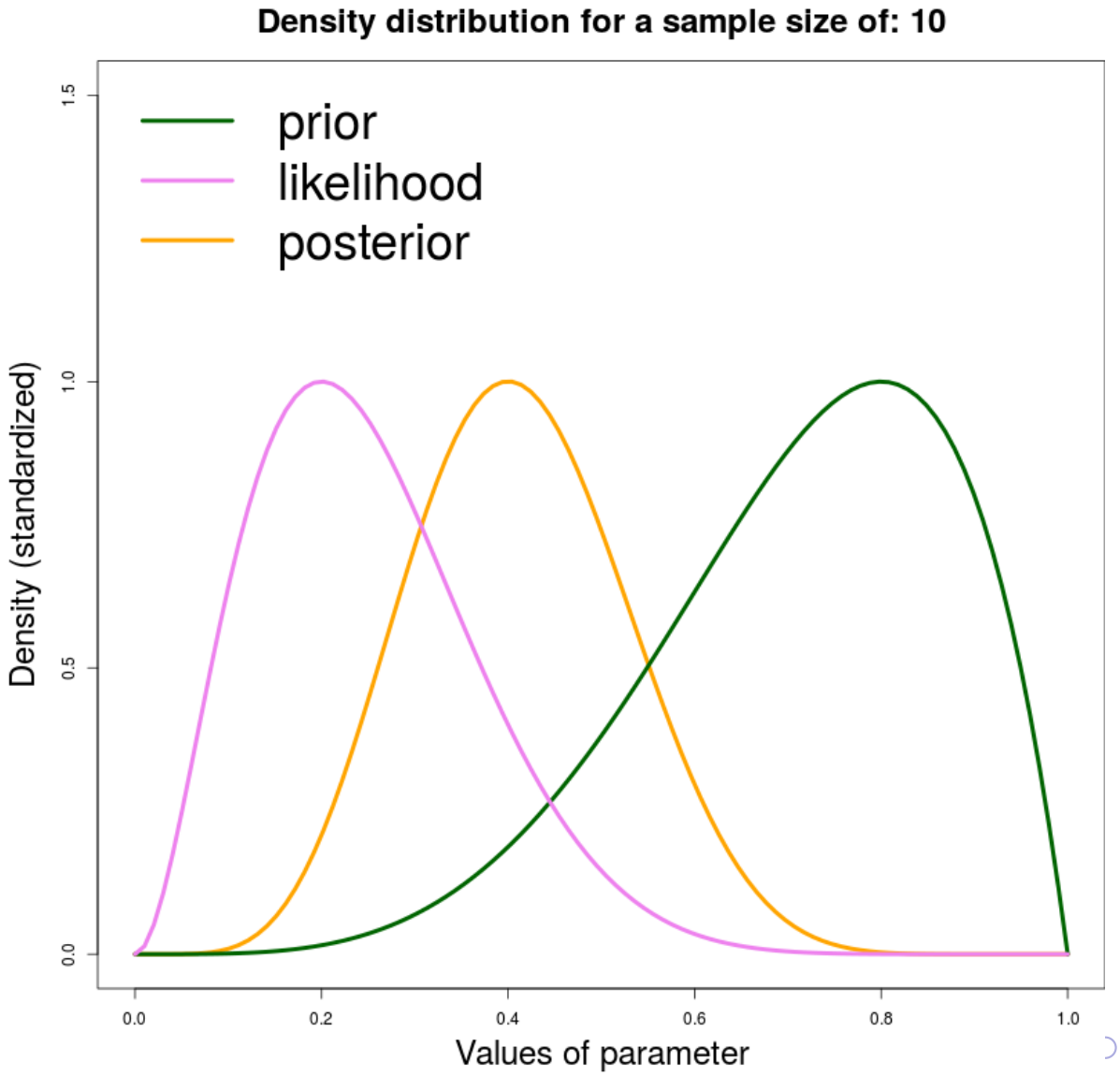
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- Weight of prior decline as sample size increases



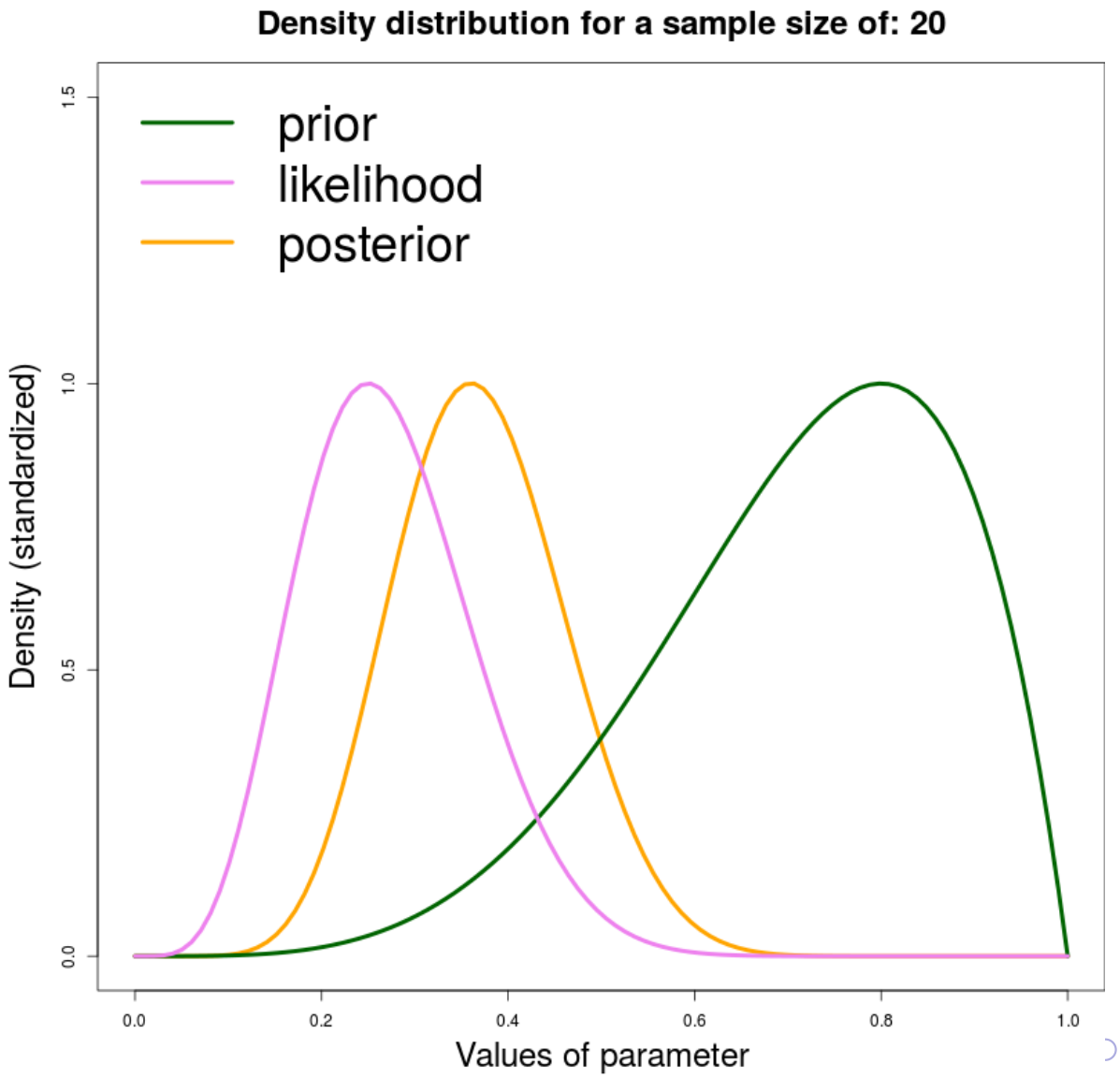
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- Weight of prior decline as sample size increases



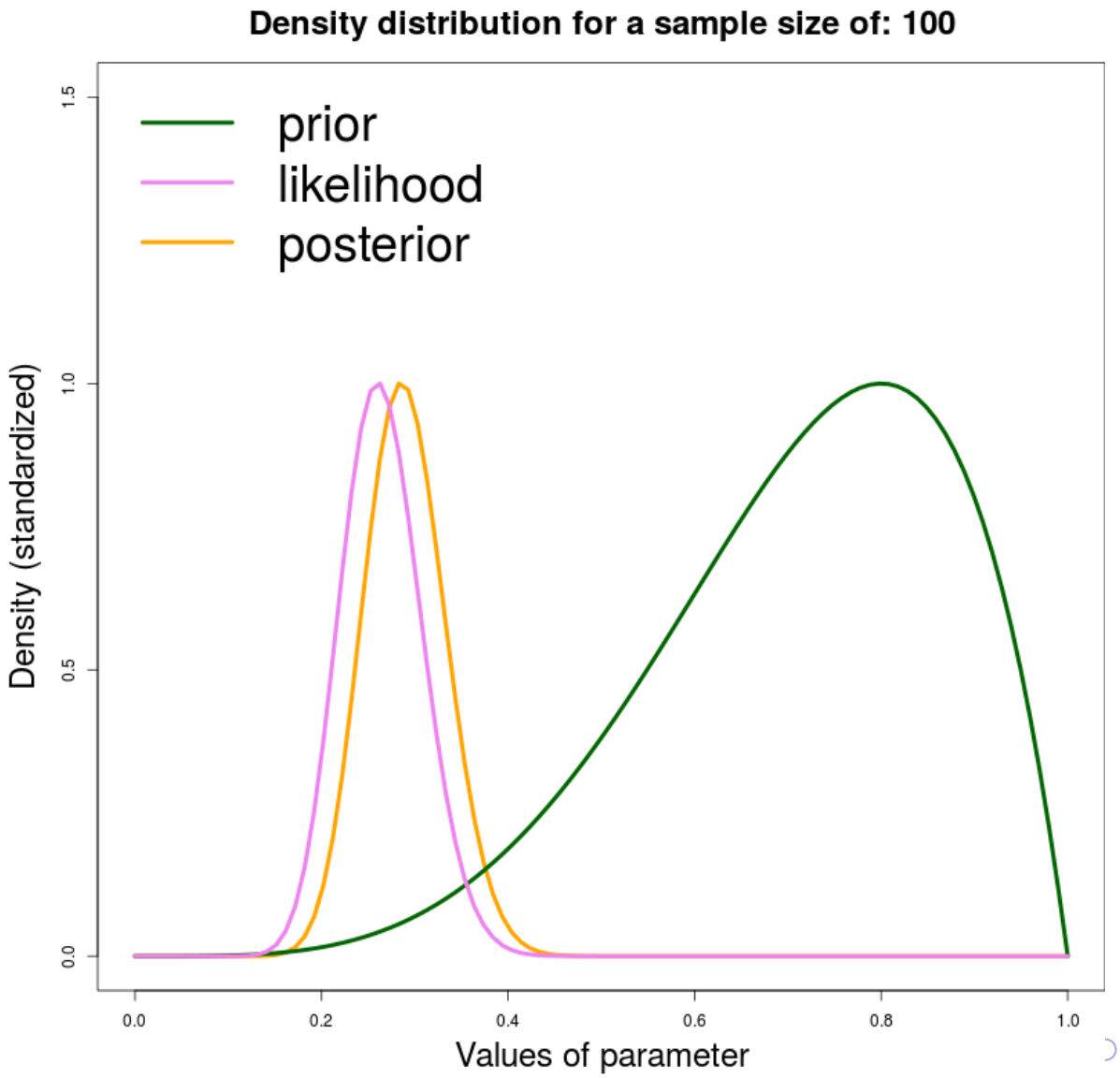
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- Weight of prior decline as sample size increases



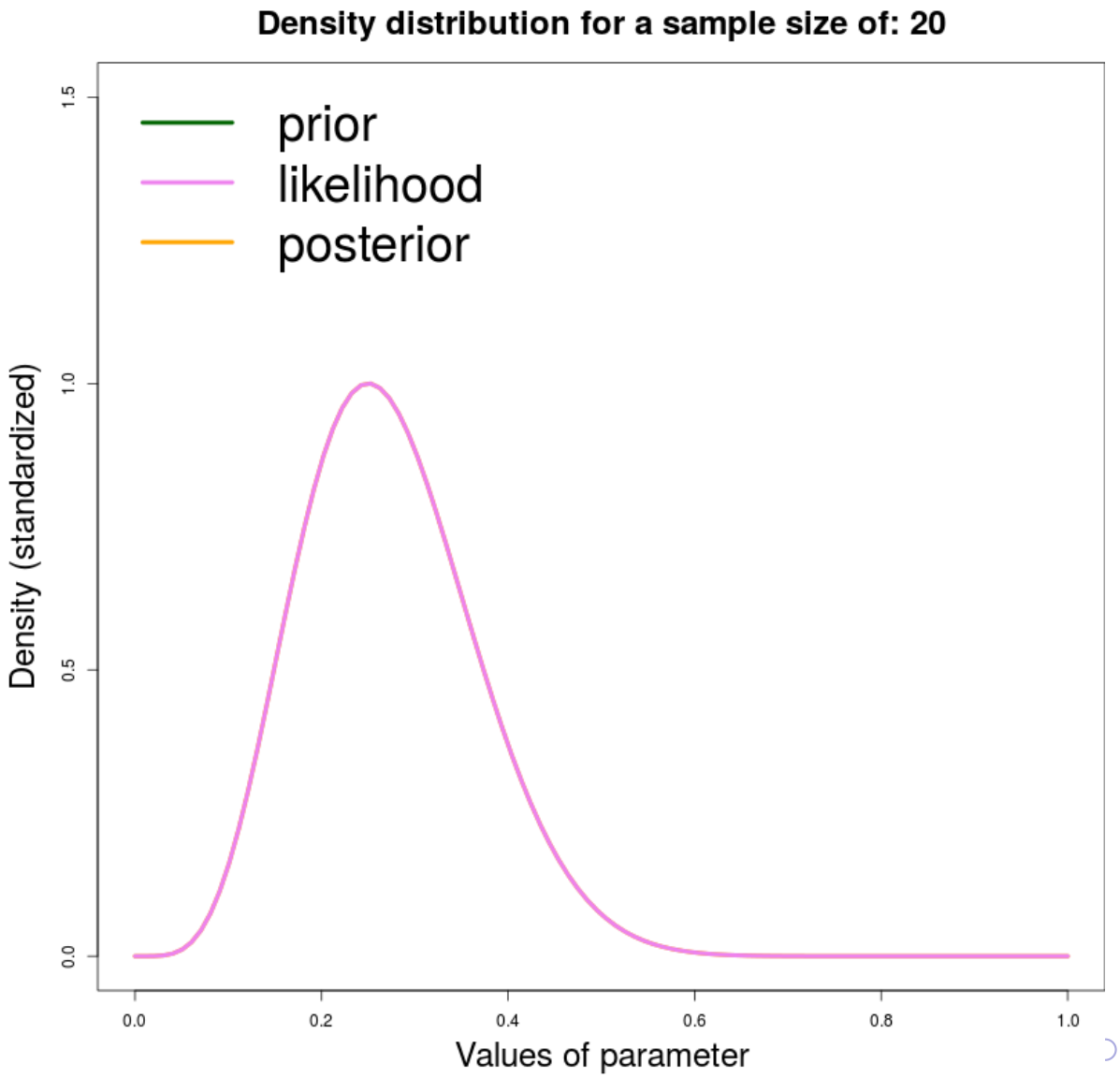
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- Weight of prior decline as sample size increases



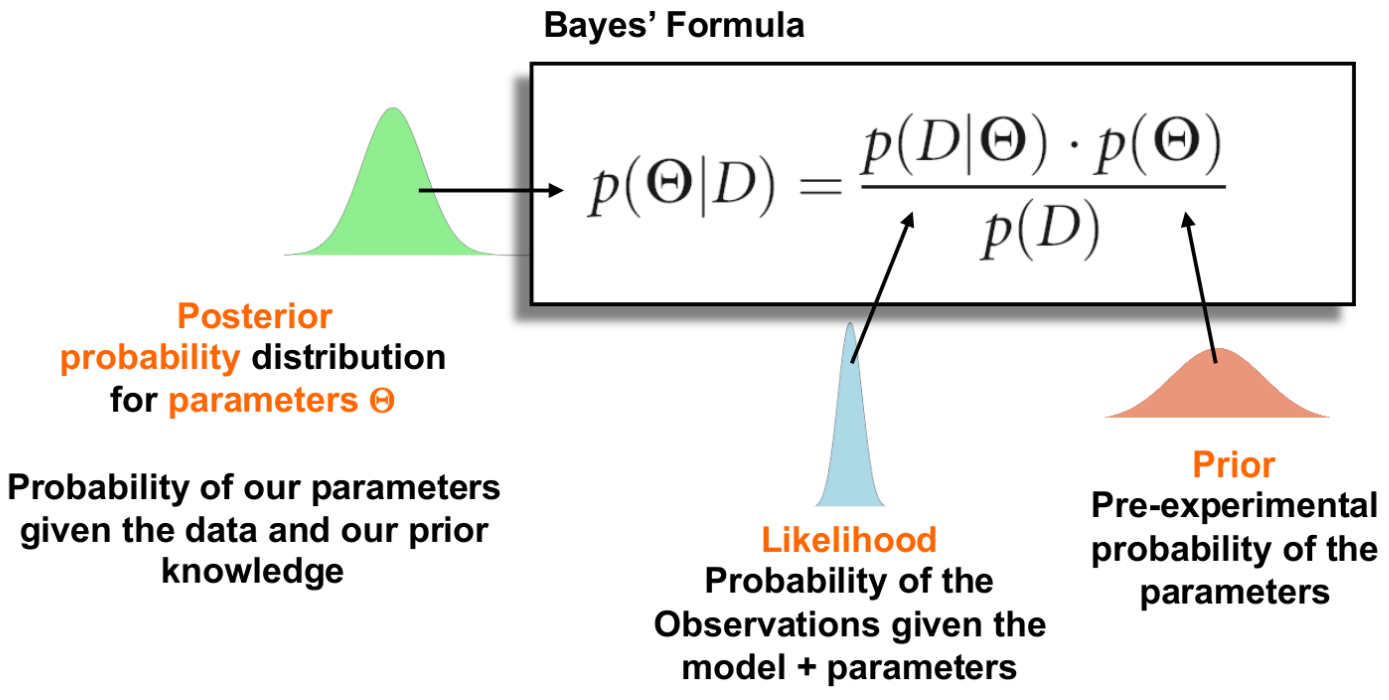
The posterior: everything we ever wanted (from a statistical model)

- Combine prior infos with new data
- Probability (density) of the parameter
- Weight of prior decline as sample size increases



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:

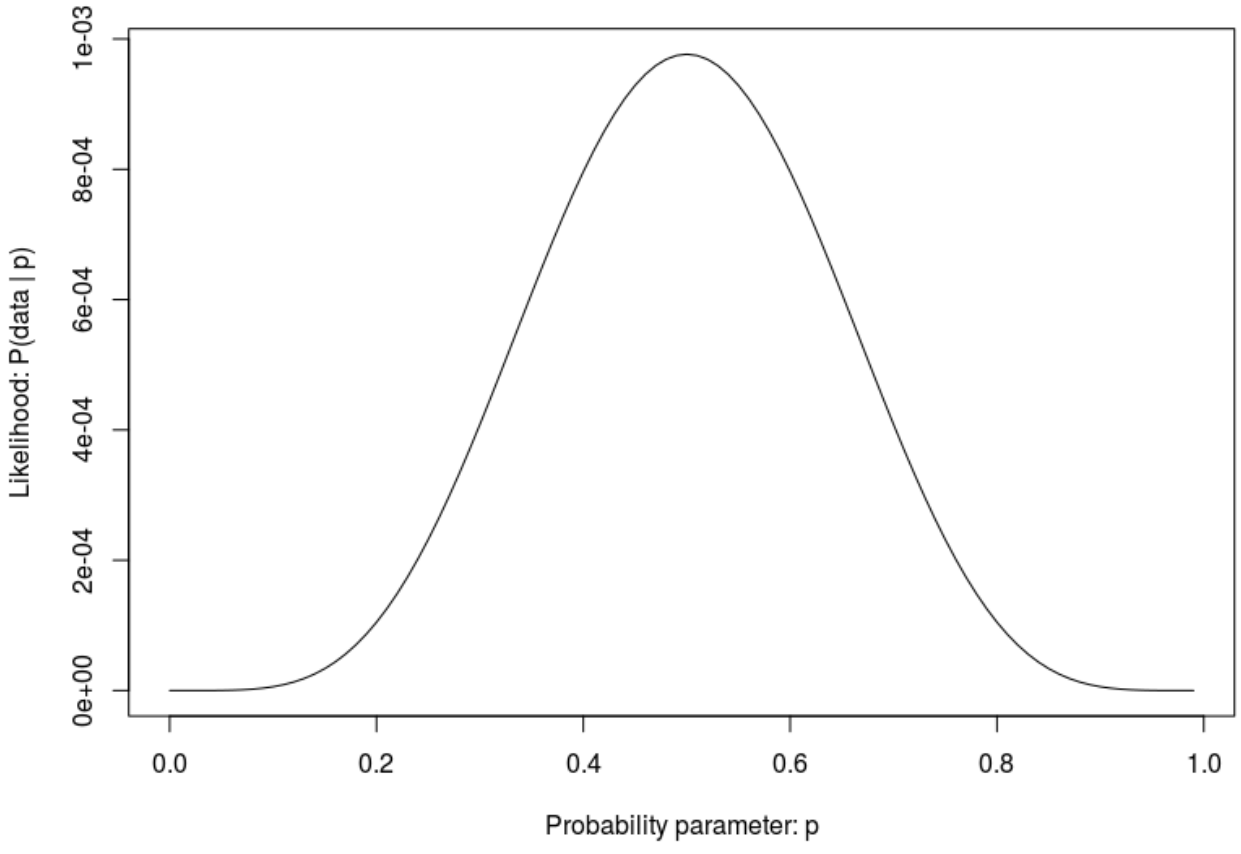
Directly using a dedicated probability language:

- JAGS via rjags
- STAN via rstan

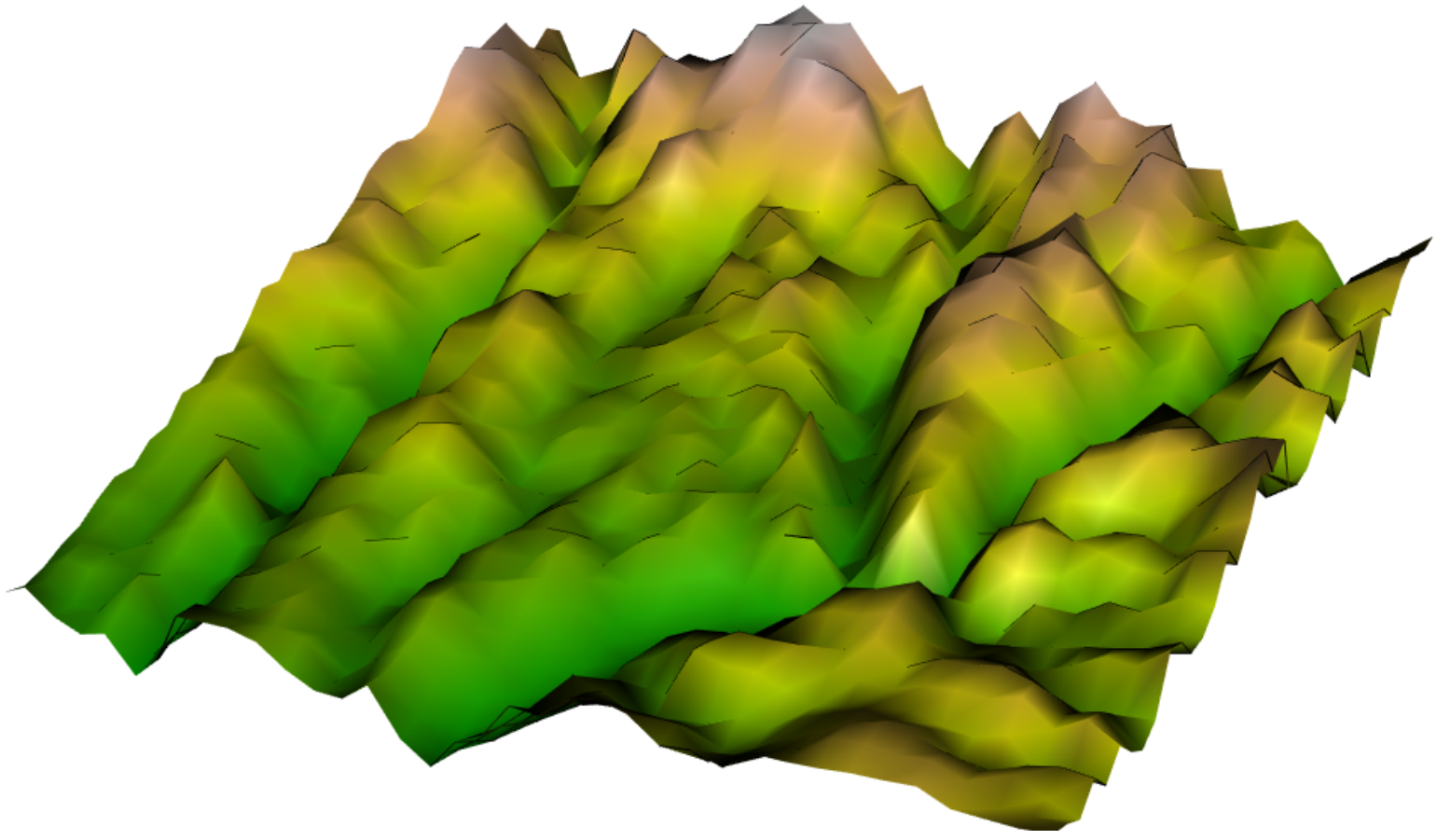
Using packages that translate R formulas into the probability language:

- rstanarm
- brms

The sampling: the issue of complex likelihood surface



The sampling: the issue of complex likelihood surface



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

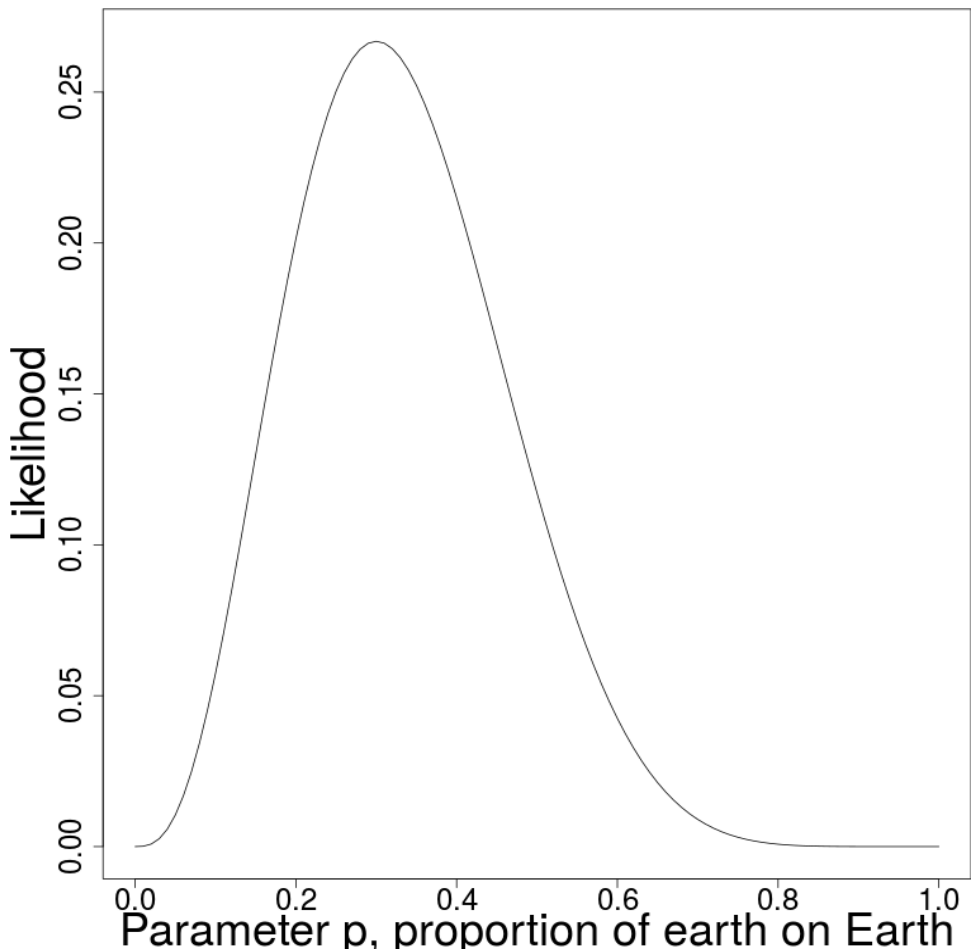
A simple Markov Chain Monte Carlo algorithm

1. Start with arbitrary parameter values and compute the posterior (likelihood times prior) at this point
2. Pick new values for all parameter in the model (based on some proposal distribution)
3. Compute the new posterior for these values
4. If the new posterior is larger than the old one, accept the new value (take one sample) if not jump to the new value based on some probability

See: `https://theoreticalecology.wordpress.com/2010/09/17/metropolis-hastings-mcmc-in-r/`

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.



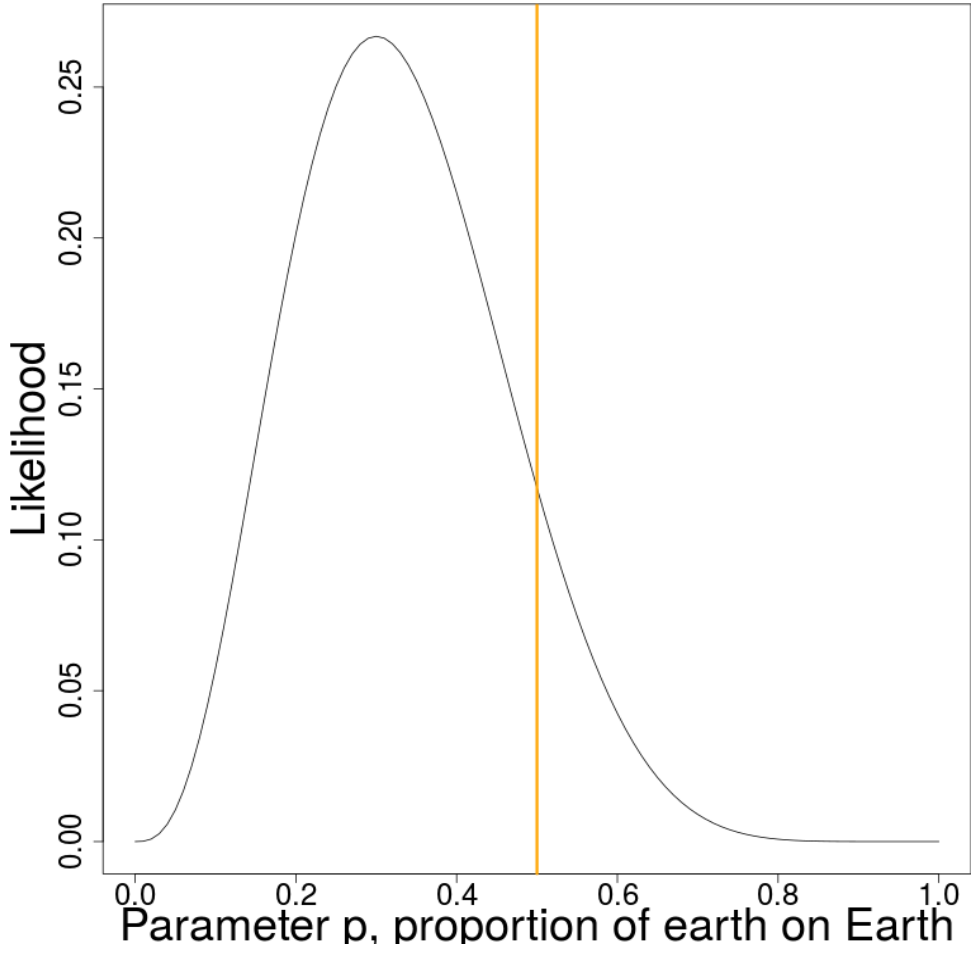
(Example from Statistical rethinking, Richard McElreath)

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

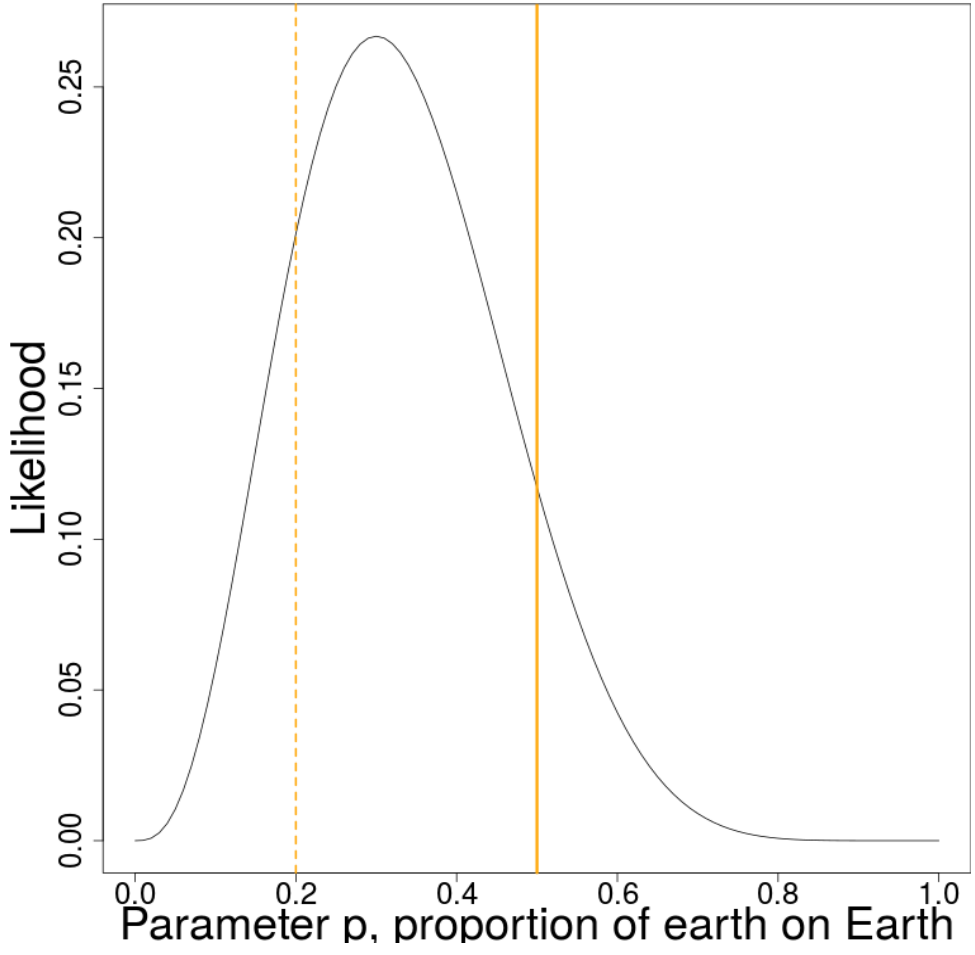


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5

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

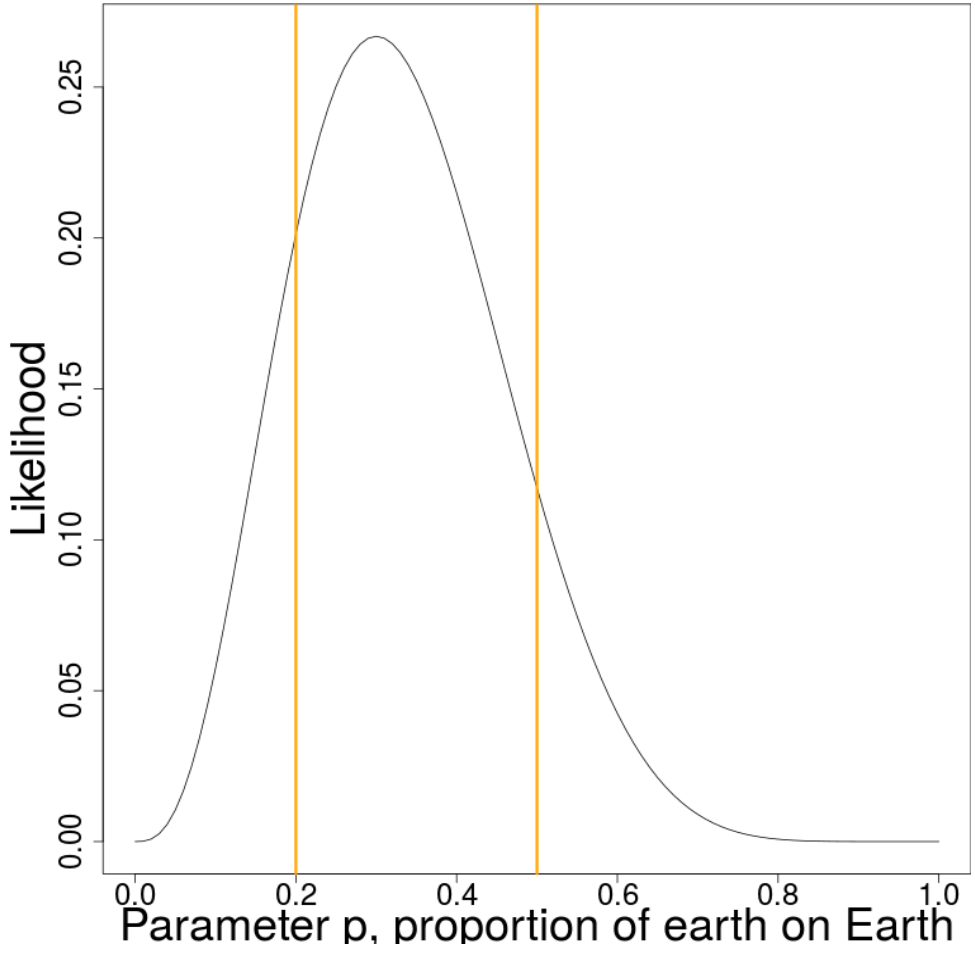


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5, 0.2

Old likelihood $<$ New likelihood, jump.

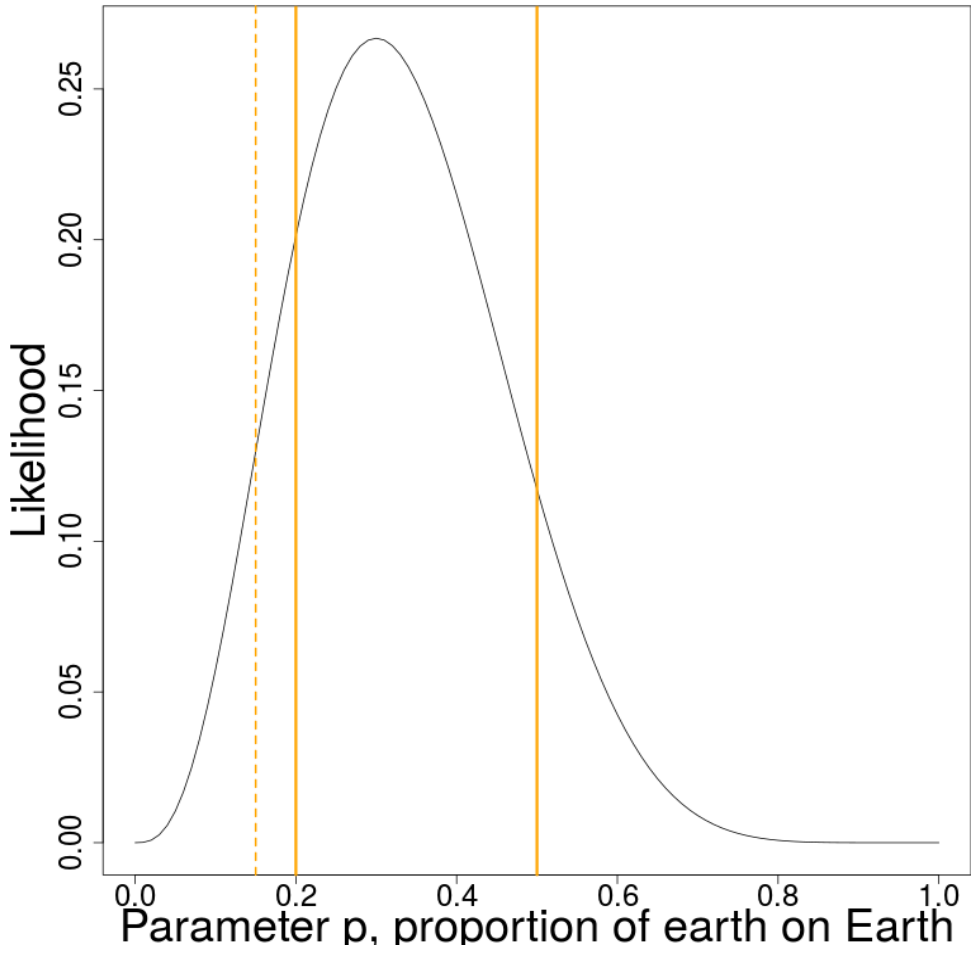


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5, 0.2

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

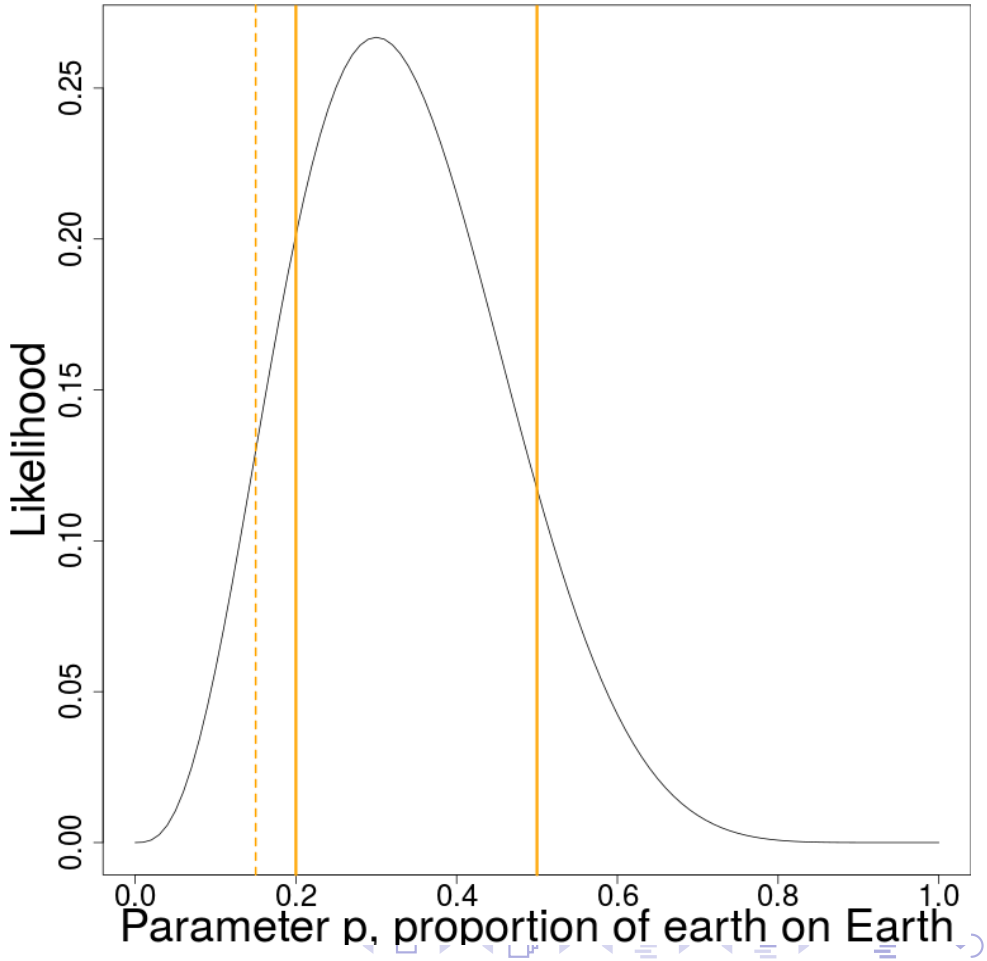


A simple MCMC example: How much earth is on Earth?

MCMC samples:

0.5, 0.2

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

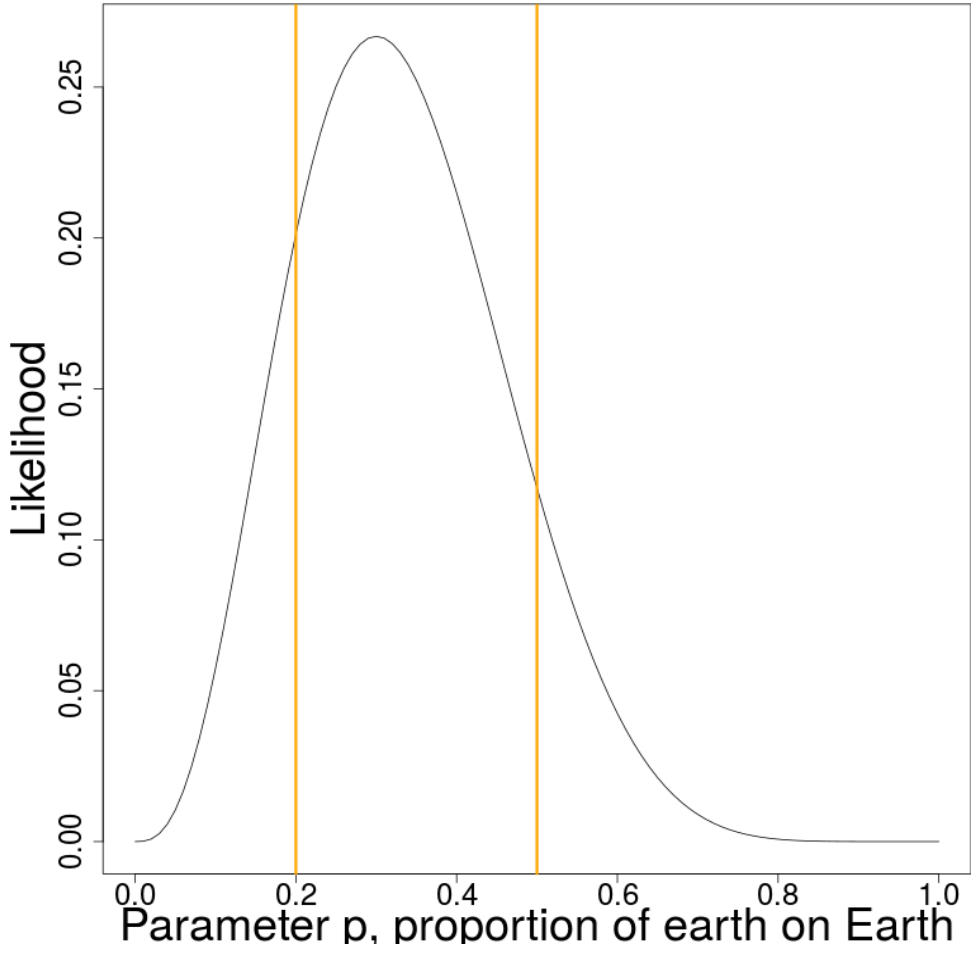


A simple MCMC example: How much earth is on Earth?

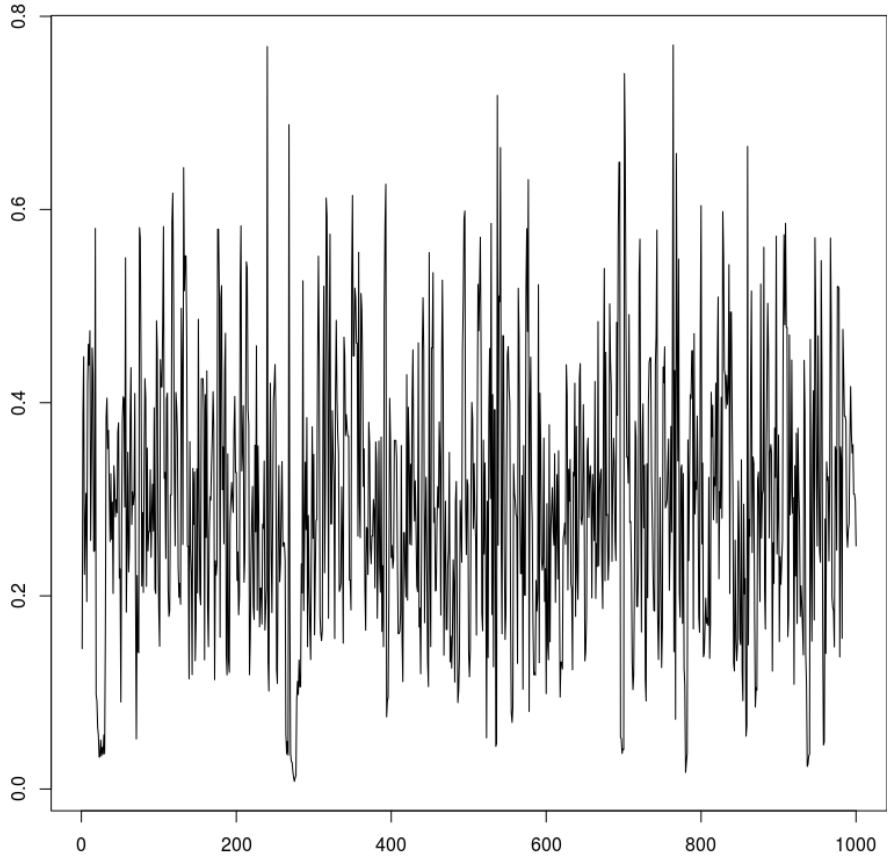
MCMC samples:

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

Different MCMC algorithm / samplers

Different MCMC algorithms are available:

- JAGS: uses (mainly) Gibbs sampling which in essence decompose a complex problem into a set of simple ones, if possible JAGS will automatically sample from the posterior (conjugacy)
- STAN: uses Hamiltonian-Monte Carlo, it gives momentum and gravity to the sampler, it is more efficient for sampling the posterior (fewer samples needed)

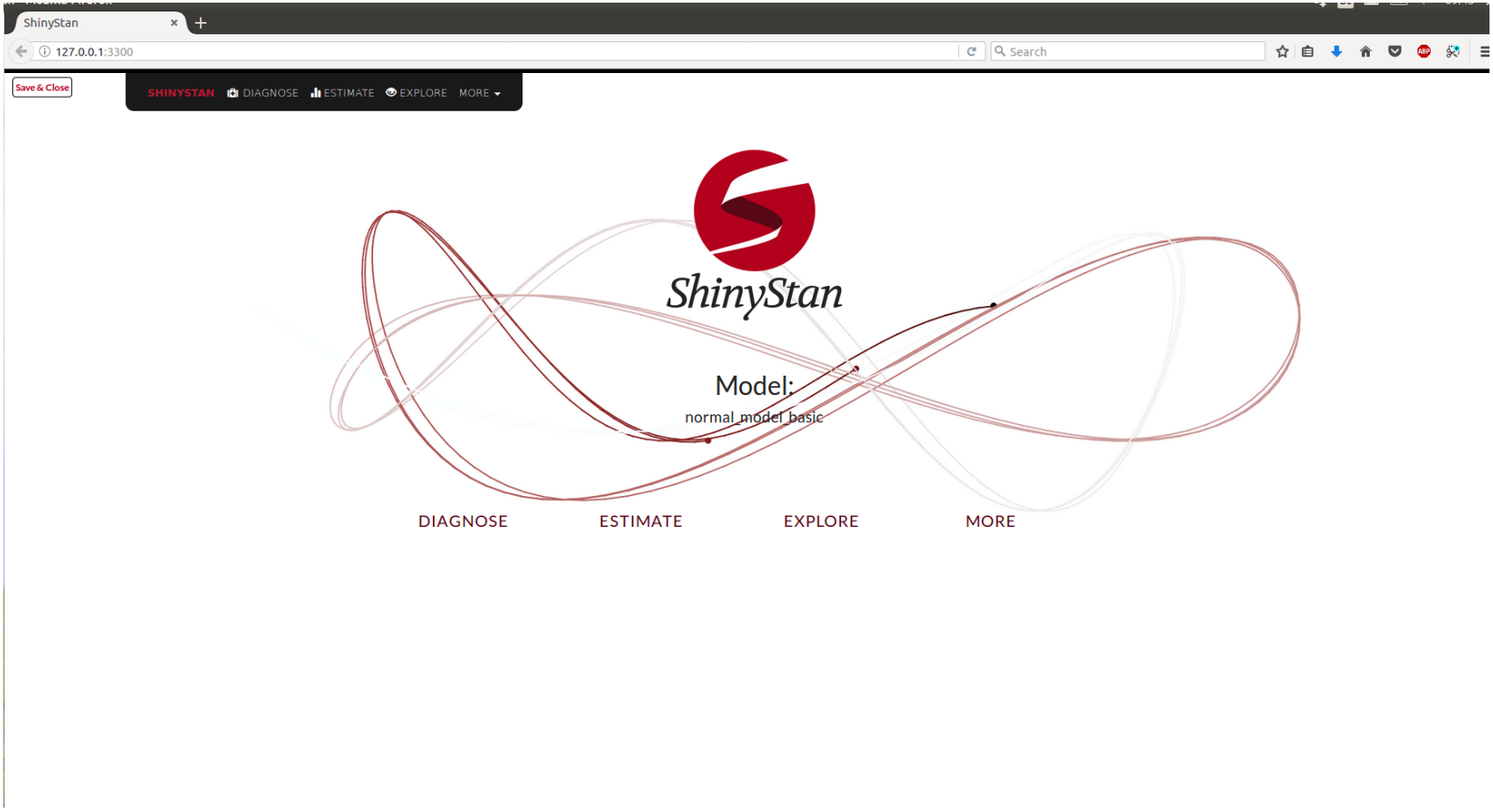
For more infos see:
M. Betancourt, `https://arxiv.org/abs/1701.02434`
C.C. Monnahan, `http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12681/full`
`https://www.youtube.com/watch?v=VnNdhsm0rJQ`

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

Your best friend for model checking is the **shinystan** package



To be explored during the coding session

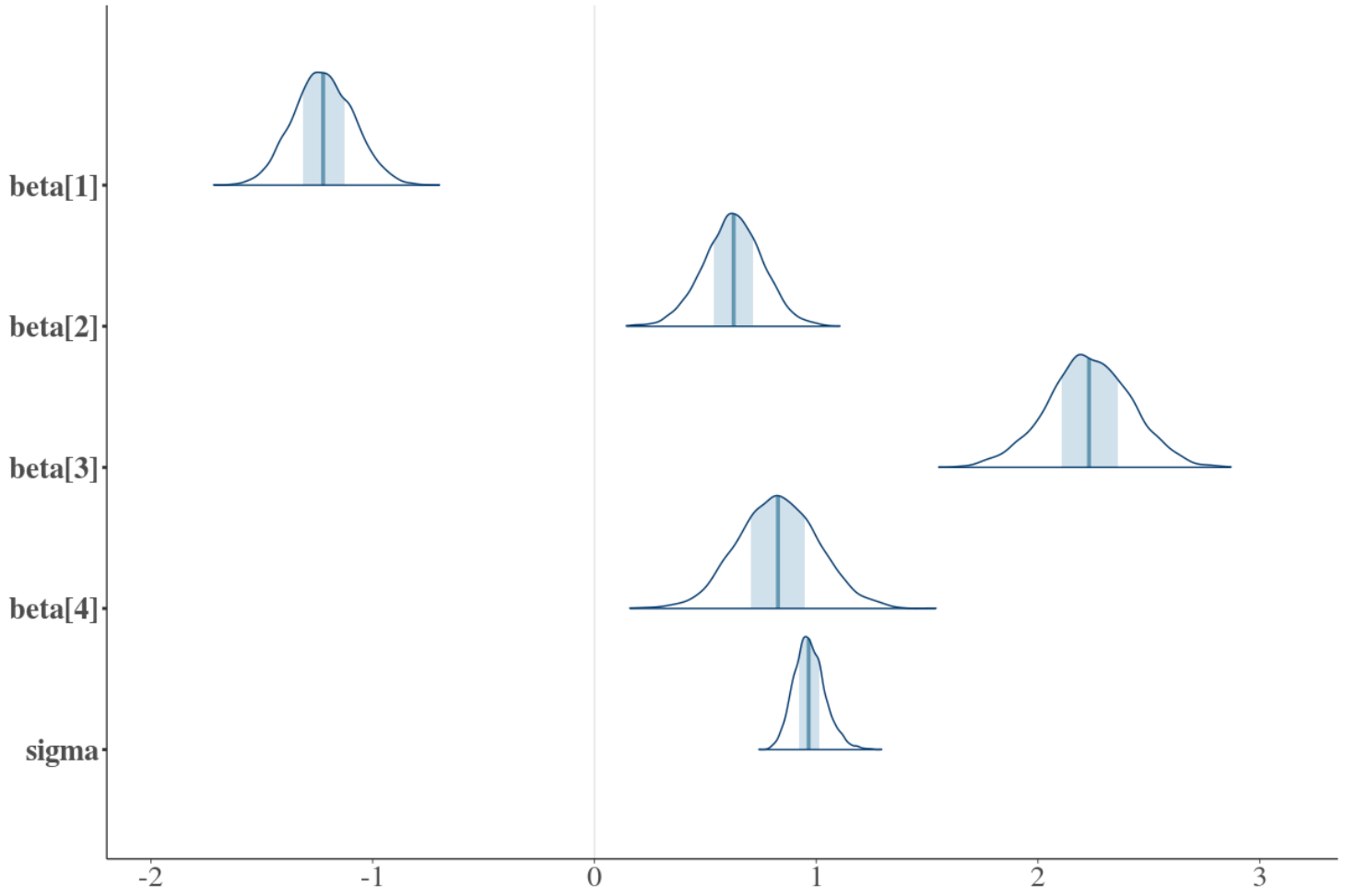
Model comparison/selection

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

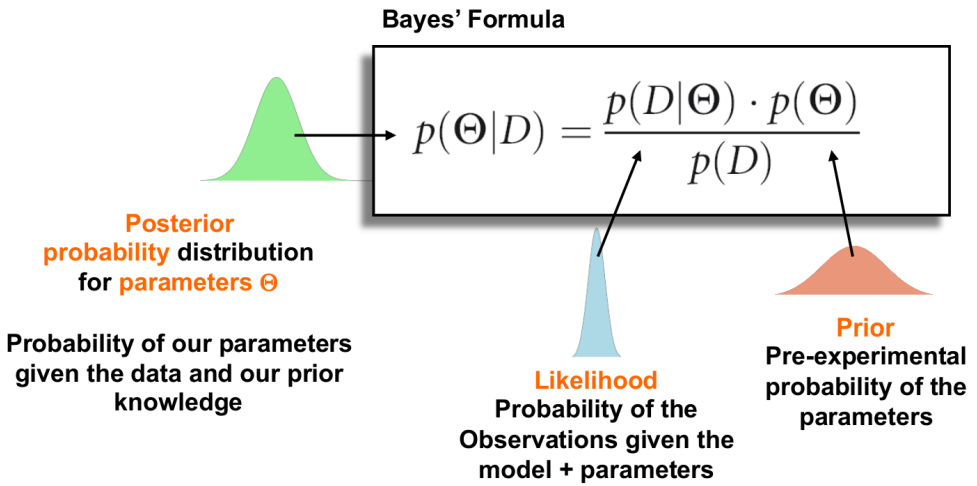
Embracing uncertainty



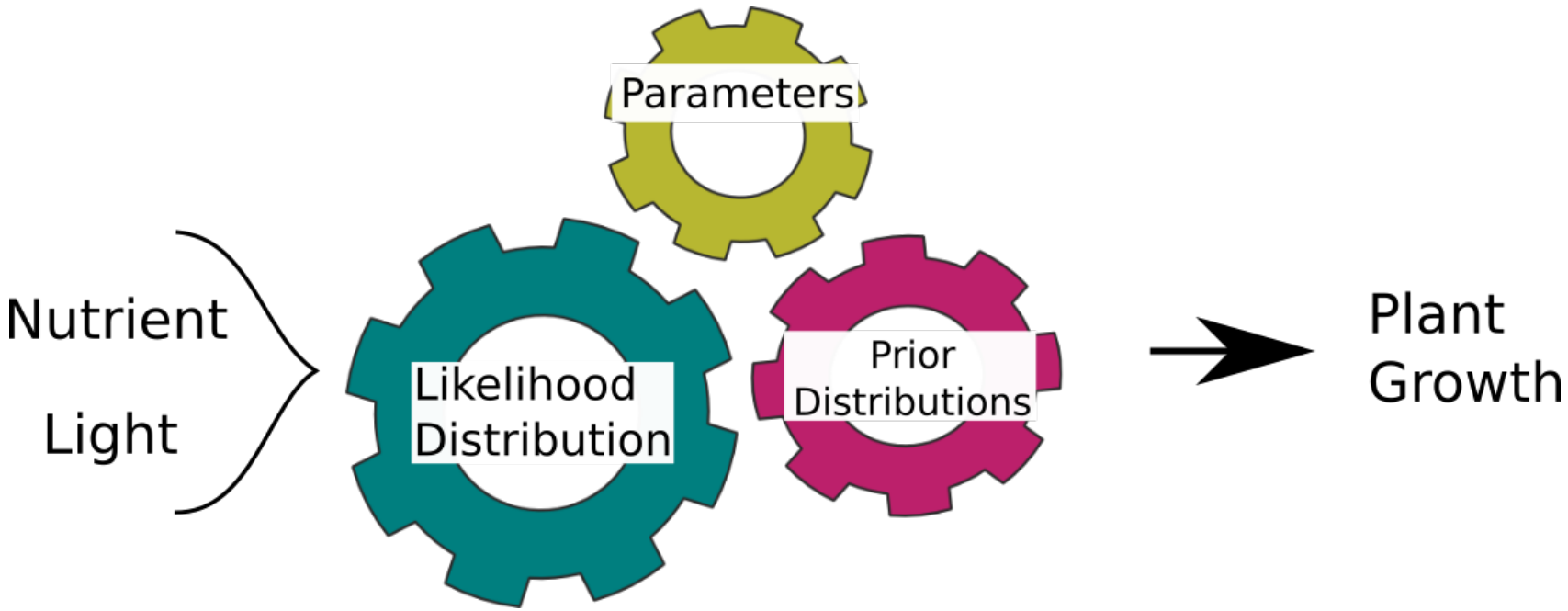
Flexibility in model building

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-generating processes



Note that this also applies to frequentist approaches (assuming fixed parameter values)

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 ...

No degrees of freedom

With Bayesian approach we can fit models with more parameters than data points

This is due to the prior distributions, it provides extra information to the model along the observed data

BUT this does not absolve us from statistical machismo (sensu Brian McGill)

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 coding

Open the *model_fitting_script.R* file