An introduction to Bayesian Data Analysis using STAN

Lionel Hertzog & Maxime Dahirel

BDA workshop, BES-GfÖ, Necov joint meeting, XX 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)

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?

Some elements of inference, how does it fits into the science workflow

Scientists are intersted with understanding and explaining the processes structuring the world, build theories to synthetically represent what is happening. Data are then assembled/collected and one wants to extract the big world signal in the data, statistical inference provide this bridge between complex data and theories.

Big world / small world

Our models have limited scope and only give answers within their area of expertise, models will tend to give the best answers within their limited scope

Bayesians can't walk on water

Except maybe shallow water

There is no bayesian model

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

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

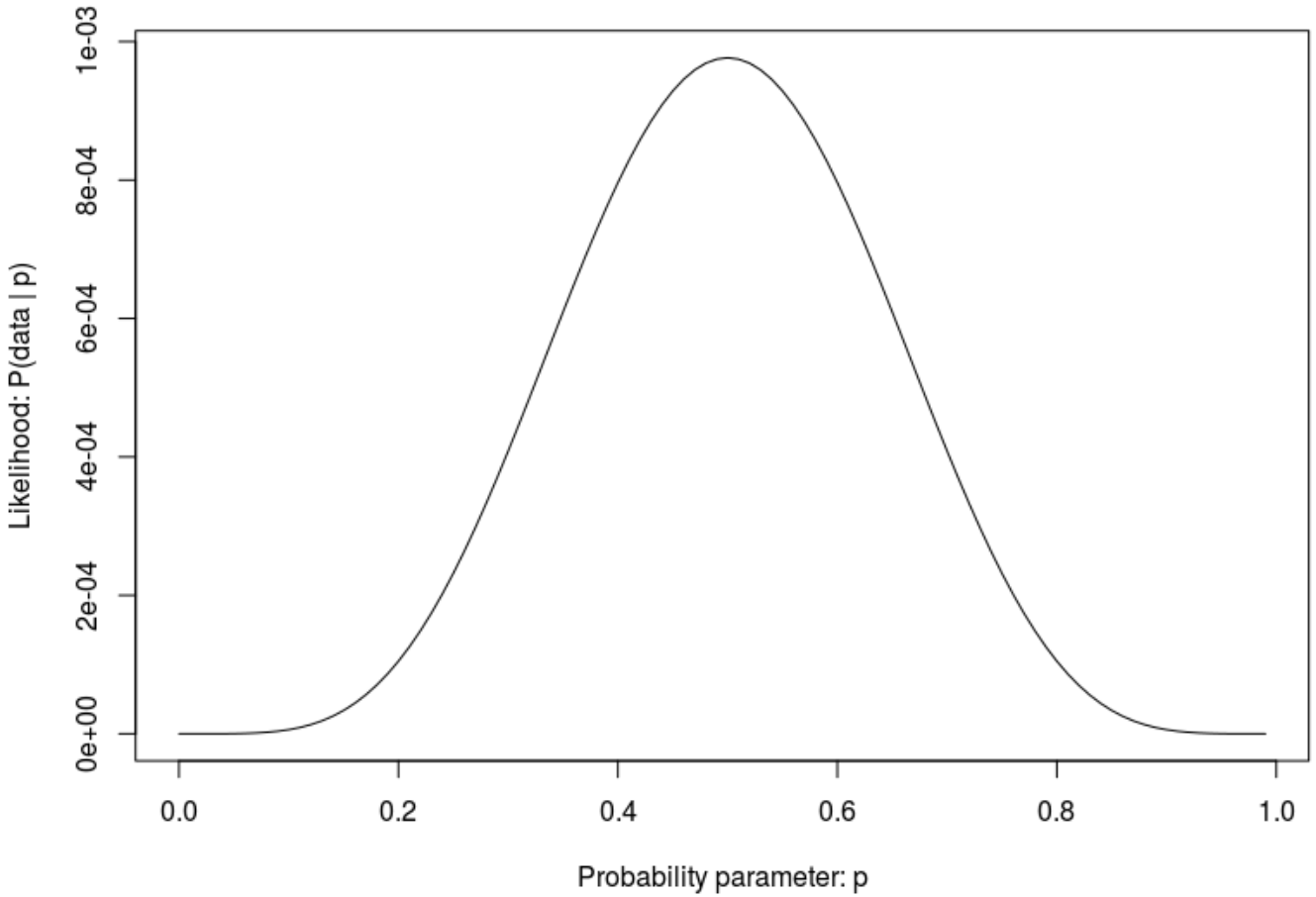
This is equivalent to:

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

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

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

The likelihood: a graphical example



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

Where BDA starts

$Posterior \propto Likelihood * Prior$

Or:

$Newknowledge \propto Currentevidence * priorknowledge$

Some graphical example of this

What is prior infos

This is something new, prior is also a probability distribution that represent our beliefs of the likely values of the parameters This is the information that is around in the literature or in the community before doing an analysis Example: want to explore plant growth or whatever with a meta-analysis

What is Posterior

The posterior is also a probability distribution, it quantifies the probability of parameter values knowing the data (the likelihood) and our expectation (prior) The impact of the data vs the expectation depends on the sample size, with low sample size the prior is relatively more important (little information in the data), as sample size grows the likelihood gets more and more weights, a simple example (maybe a shiny app ...)

The key aspects of BDA

Explicitely integrate prior knowledge Output probability for the parameters/hypothesis, posterior Uncertainty is everywhere Think of models as data-generating processes

Ways to fit bayesian models in R

Coding in probability language vs using wrappers STAN is a programming language in its own so can code the models direclty in it (some snippet code example of simple model) Knowing this we have the full flexibility to fit any model we want but we also have the full possibility that we make coding/interpretation errors that are not visible. There are a couple of R packages to transcribe the standard R formula synthax into STAN models: rstanarm and brms. With this option we are sure that the model code is correct and also optimized so it will certainly run faster than naive implementation directly in STAN. But one is limited by what the package developped have implemented.

The sampling

Blind man in the likelihood landscape, how do we effectively sample it

About MCMC

Markov Chain Monte Carlo: stochastic transition within parameter space only based on current and proposed jump.

Different samplers

JAGS: Adaptively tries to find the best sampler corresponding to your model STAN: Use Hamiltonian Monte Carlo (see Monnahan et al 2017 MEE)

Elements of Bayesian vocabulary

Chains, convergence, divergence, Rhat, n_eff

Model checking in Bayesian data analysis

convergence, n_eff, chains, posterior predictive checks

Model comparison/selection

LOOC, WAIC

Embracing uncertainty

Everything is variable, all parameters come with (posterior) distribution, can interprete posterior sample as probabilities, flexibility to test any hypothesis building on these

Flexibility in model building

As long as you can write your likelihood function you can fit any model you like (similar to using MLE approach ie with bbmle)

Why do we do stats?

Bayesian data analysis give us relevant answers in terms of probability rather than weird answers refering to som null hypothesis in terms of frequency

No degrees of freedom

Fit complex models even with little data (is this something we want??)

Asymptotic convergence, when bayesian and frequentist approach give similar answers

With infinite sample size the posterior distribution just reflect the likelihood, as sample size increases priors have dwindling effects