# Session 1.1: Bayesian thinking

Spatial and Spatio-Temporal Bayesian Models with R-INLA, University of São Paulo

26 September 2022

# Learning objectives

After this lecture you should be able to

- Introduce Bayesian way of thinking
- Present Bayes theorem and Bayesian inference
- Describe computational methods commonly used to perform Bayesian inference

The topics treated in this lecture are presented in Chapter 3-4 of the book **Spatial and Spatio-Temporal Bayesian models with R-INLA**.

# Outline

1. Why Bayesian

2. Components of a Bayesian analysis

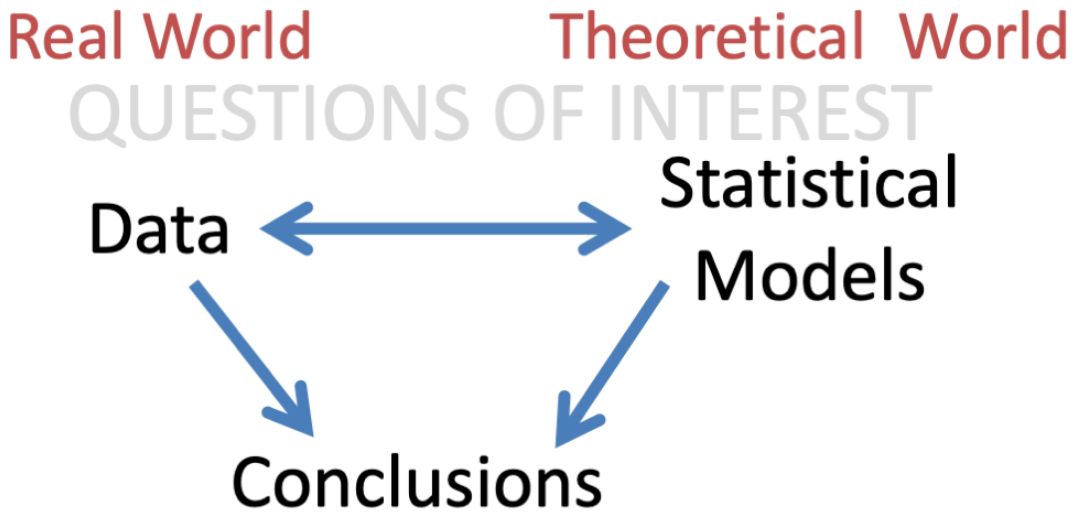
3. Bayes Theorem

4. Bayesian computing

# Why Bayesian

QUESTIONS OF INTEREST

# Statistics - The 'Big Picture'




- Several different ways of formulating statistical models and computing inferences from these models and data
- Can be grouped into two broad approaches:
  – Frequentist
  – Bayesian

# Everyday thought process

- At the start of the football season I formed a view about the chance that the team I support will be relegated (**PRIOR**), based on
  - performance last season
  - summer transfers
  - etc.

# Everyday thought process

- At the start of the football season I formed a view about the chance that the team I support will be relegated (**PRIOR**), based on
  - performance last season
  - summer transfers
  - etc.
- The first match is played
- Now I have some information from current season (**DATA**)

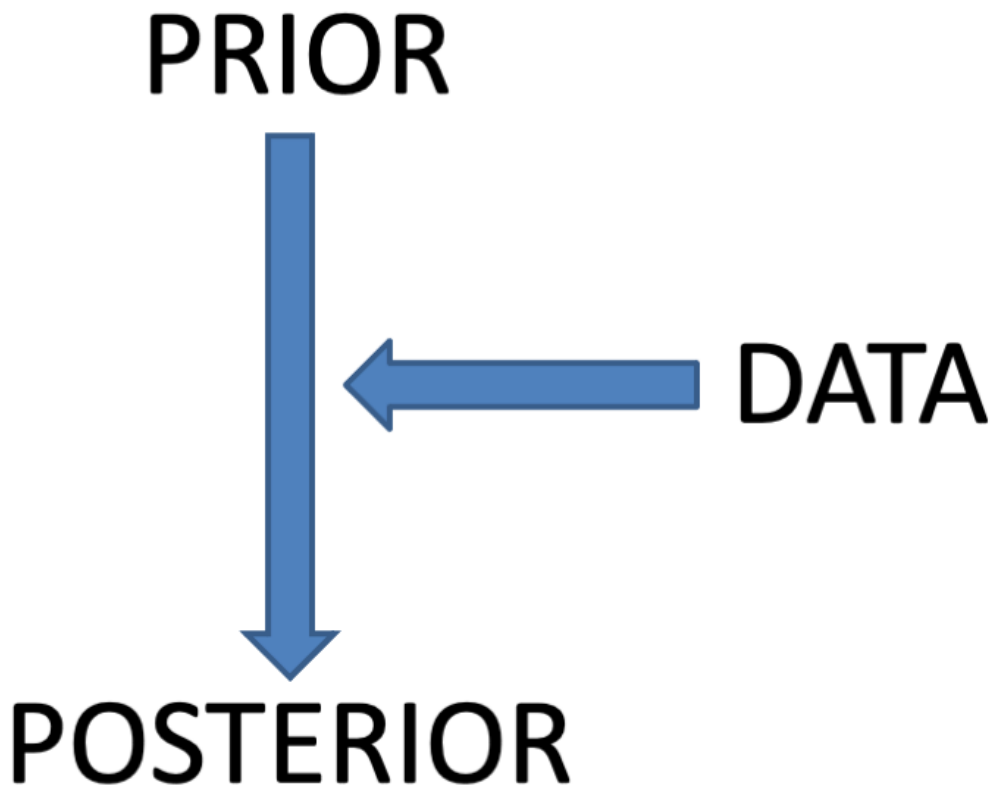
# Everyday thought process

- At the start of the football season I formed a view about the chance that the team I support will be relegated (**PRIOR**), based on
  – performance last season
  – summer transfers
  – etc.
- The first match is played
- Now I have some information from current season (**DATA**)
- I re-assesses the probability of relegation upwards, because
  – they lose
  – their main striker limps off
- Prior view updated with data to give **POSTERIOR** view

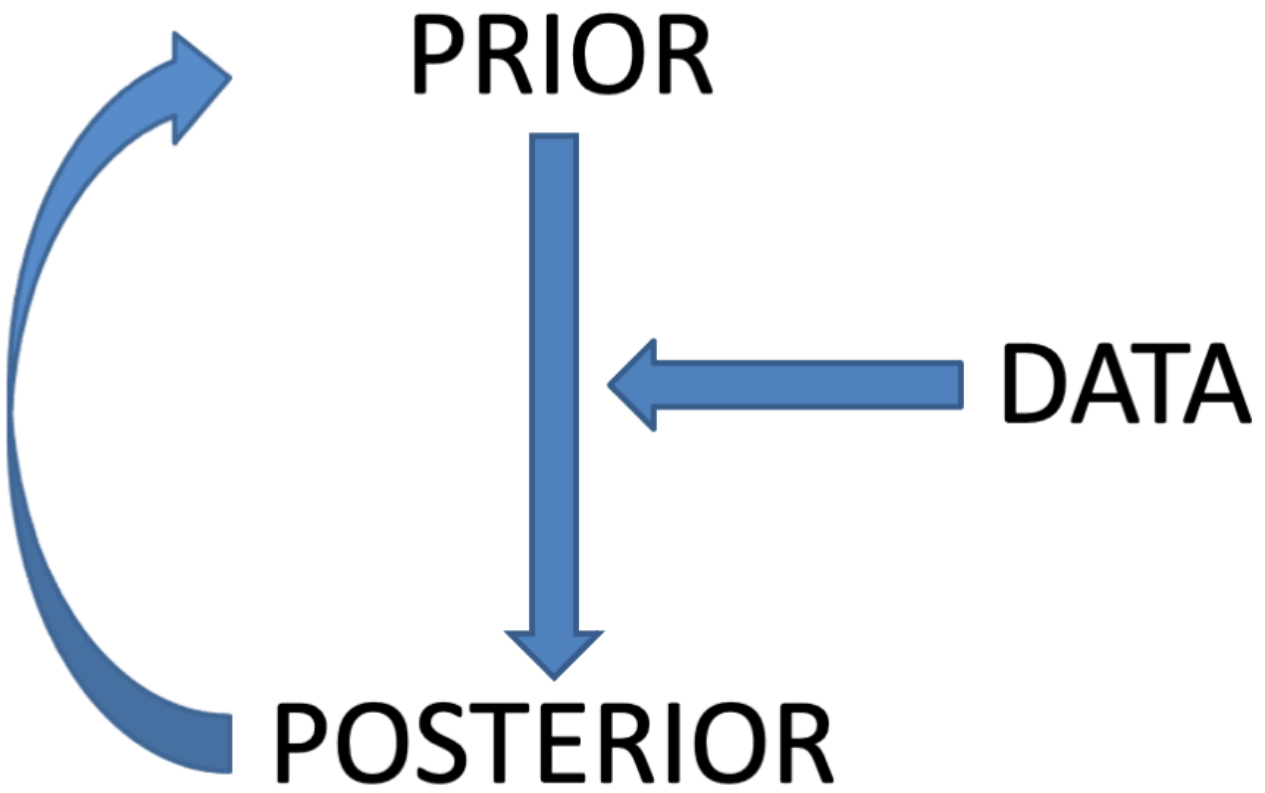
# Thought process of a physician

- A patient presents with a set of symptoms, concerned that they might have a certain disease
- The physician assesses the chance that the patient has this disease, based on
  - symptoms
  - family history
  - alternative explanations of symptoms
  - prevalence of disease
- The physician sends the patient for a diagnostic test
- The physician re-assesses the chance that the patient has this disease, taking account of
  - results of diagnostic test
  - reliability of diagnostic test
- The physician may send the patient for further diagnostic tests

# Bayesian thinking



# Bayesian thinking



# Why Bayesian methods?

- Bayesian methods have been widely applied in many areas:
  - medicine / epidemiology
  - genetics
  - ecology
  - environmental sciences
  - social and political sciences
  - finance
  - archaeology
  - .....
- Motivations for adopting Bayesian approach vary:
  - natural and coherent way of thinking about science and learning
  - pragmatic choice that is suitable for the problem in hand

# Example

A clinical trial is carried out to collect evidence about an unknown *treatment effect*

Conventional analysis

- p-value for $H_0$: treatment effect is zero
- Point estimate and CI as summaries of size of treatment effect

**Aim is to learn what this trial tells us about the treatment effect**

# Example

A clinical trial is carried out to collect evidence about an unknown *treatment effect*

Conventional analysis

- p-value for $H_0$: treatment effect is zero
- Point estimate and CI as summaries of size of treatment effect

**Aim is to learn what this trial tells us about the treatment effect**

Bayesian analysis

- Inference is based on probability statements summarising the posterior distribution of the treatment effect

Asks: **how should this trial change our opinion about the treatment effect?**

# Components of a Bayesian analysis

# Components of a Bayesian analysis

A clinical trial is carried out to collect evidence about an unknown *treatment effect*. The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the treatment effect *excluding* the evidence from the trial (the prior distribution)
- the support for different values of the treatment effect based *solely* on data from the trial (the likelihood),

and to combine these two sources to produce

- a final opinion about the treatment effect (the posterior distribution)

# Components of a Bayesian analysis

A clinical trial is carried out to collect evidence about an unknown *treatment effect*. The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the treatment effect *excluding* the evidence from the trial (the prior distribution)
- the support for different values of the treatment effect based *solely* on data from the trial (the likelihood),

and to combine these two sources to produce

- a final opinion about the treatment effect (the posterior distribution)

The final combination is done using **Bayes theorem** (and only simple rules of probability), which essentially weights the likelihood from the trial with the relative plausibilities defined by the prior distribution

One can view the Bayesian approach as a formalisation of the process of learning from experience

# Bayesian inference: the posterior distribution

Posterior distribution forms basis for all inference --- can be summarised to provide

- point and interval estimates of Quantities of Interest (QOI), e.g. treatment effect, small area estimates, ...
- point and interval estimates of any function of the parameters
- probability that QOI (e.g. treatment effect) exceeds a critical threshold
- prediction of QOI in a new unit
- prior information for future experiments, trials, surveys, ...
- inputs for decision making
- ...

# Bayes theorem and its link with Bayesian inference

**Bayes' theorem**

- Provable from probability axioms
- Let $A$ and $B$ be events, then

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

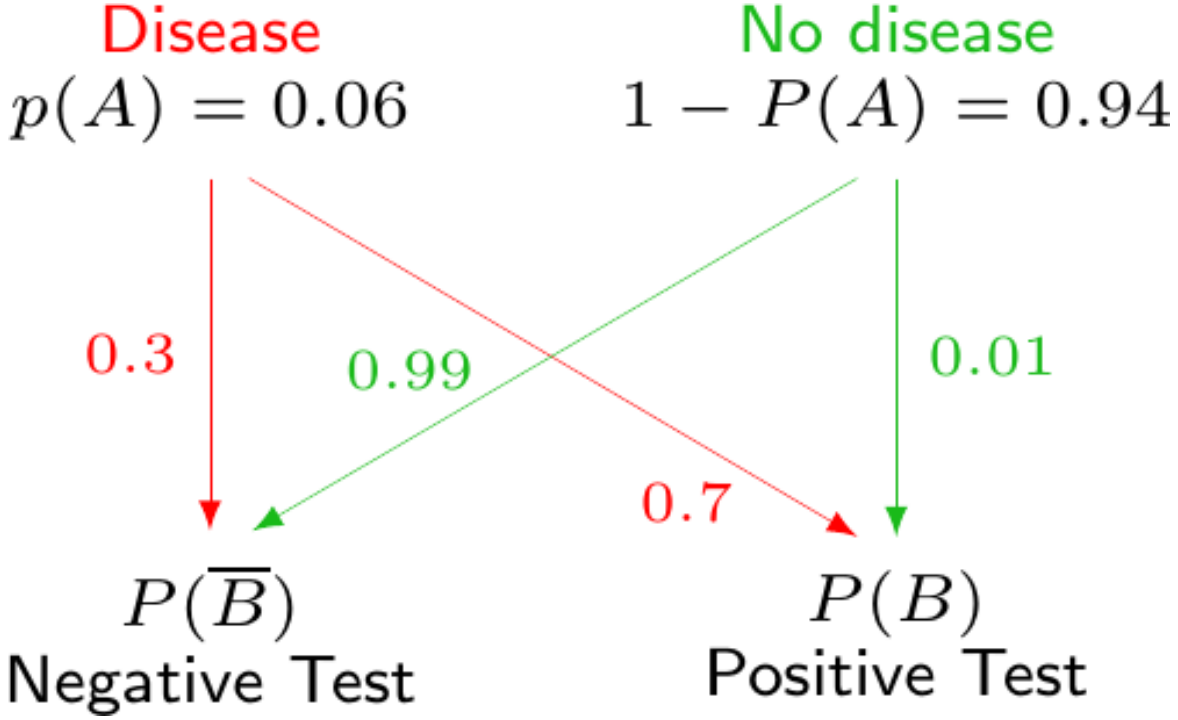
- If $A_i$ is a set of mutually exclusive and exhaustive events (*i.e.* $A_i \cap A_j = \emptyset, p(\bigcup_i A_i) = \sum_i p(A_i) = 1$), then

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{p(B)} = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)}$$

# An example: diagnostic tests

The latest COVID-19 test has shown to have 70% sensitivity and 99% specificity

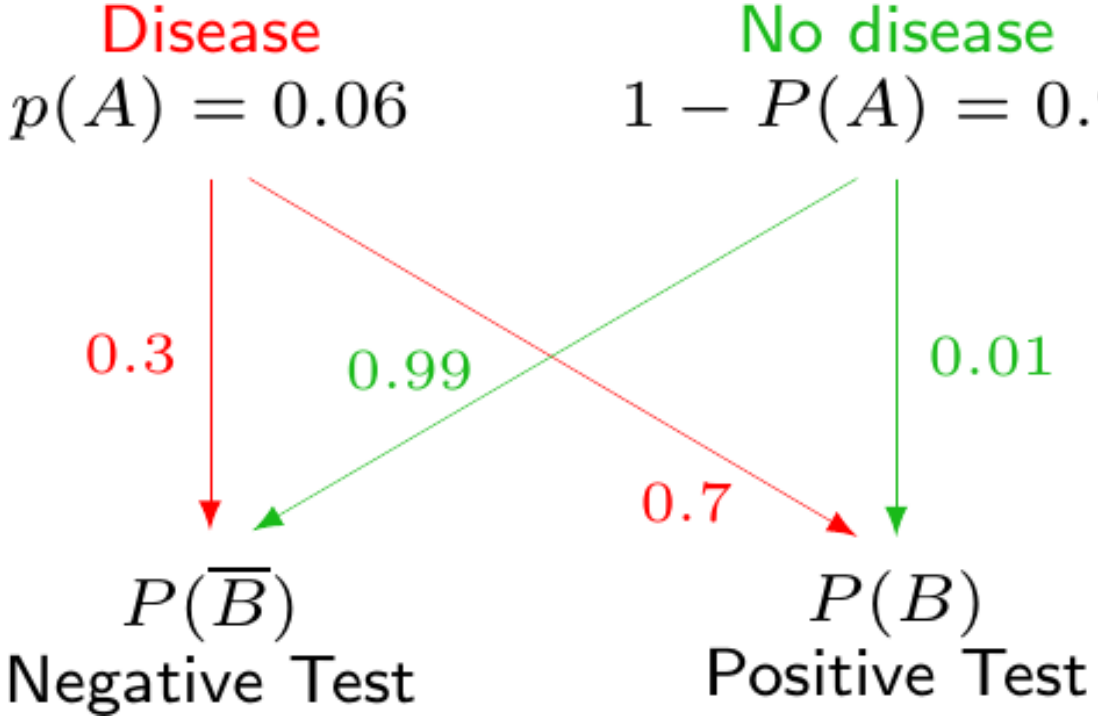
In England, COVID prevalence is 6%; what is the chance that a patient testing positive actually does have COVID-19?



# An example: diagnostic tests

The latest COVID-19 test has shown to have 70% sensitivity and 99% specificity

In England, COVID prevalence is 6%; what is the chance that a patient testing positive actually does have COVID-19?



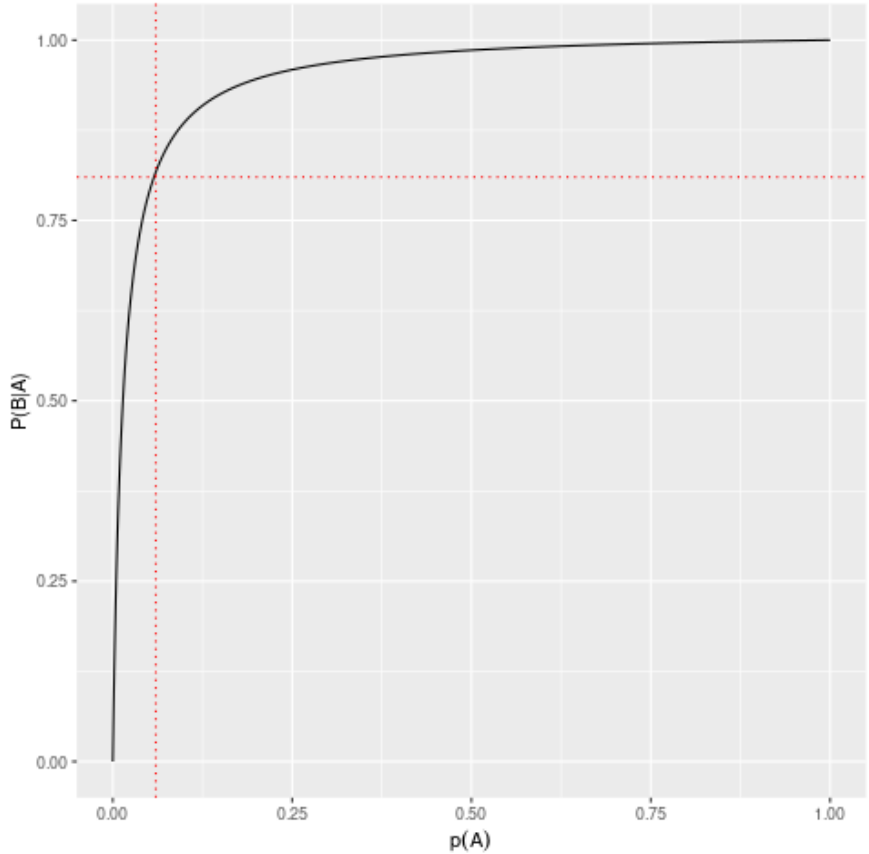

We are interested in

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\overline{A})p(\overline{A})} = \frac{0.7 \times 0.06}{0.7 \times 0.06 + 0.01 \times 0.94} = 0.81$$

# Comments

- The disease prevalence can be thought of as a *prior* probability ( $p$ = 0.06)

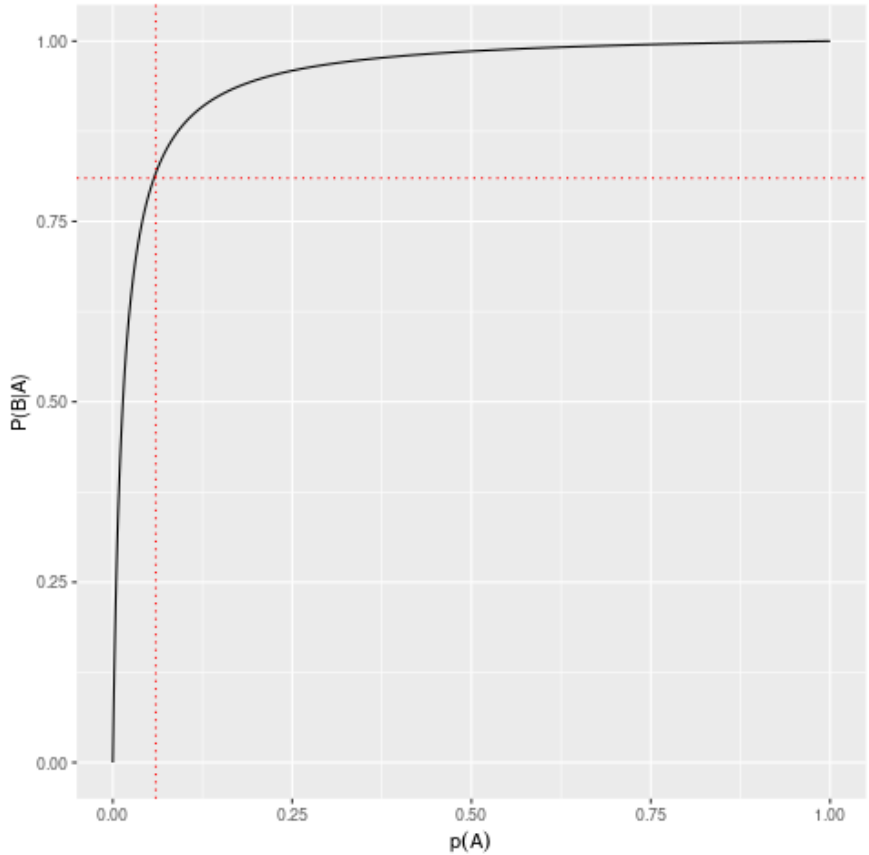
- Observing a positive result causes us to modify this probability to $p$ = 0.81. This is our *posterior* probability that patient is COVID-19 positive.



# Comments

- The disease prevalence can be thought of as a *prior* probability ( $p$ = 0.06)

- Observing a positive result causes us to modify this probability to $p$ = 0.81. This is our *posterior* probability that patient is COVID-19 positive.




- Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established
- More controversial in general statistical analyses: *parameters* are unknown quantities, and prior distributions need to be specified $\rightarrow$ Bayesian inference

# Bayesian inference

Makes fundamental distinction between

- Observable quantities $y$, i.e. the data
- Unknown quantities $\theta$

  $\theta$ can be statistical parameters, missing data, mismeasured data...

  $\rightarrow$ parameters are treated as random variables

  $\rightarrow$ in the Bayesian framework, we make probability statements about model parameters

in the frequentist framework, parameters are fixed non-random quantities and the probability statements concerning the data

As with any statistical analysis, we start building a model which specifies $p(y \mid \theta)$

This is the likelihood, which relates all variables into a full probability model

# Bayesian inference [continued]

From a Bayesian point of view

- $\theta$ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data

$\rightarrow$ need to specify a prior distribution $p(\theta)$

- $y$ is known so we should condition on it

$\rightarrow$ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, d\theta} \propto p(\theta)\, p(y \mid \theta)$$

- This is the posterior distribution

The prior distribution $p(\theta)$, expresses our uncertainty about $\theta$ before seeing the data

The posterior distribution $p(\theta \mid y)$, expresses our uncertainty about $\theta$ after seeing the data

# Bayesian computation

# How to obtain the posterior distribution?

- When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood $\rightarrow$ the posterior is a known distribution.
- In real life it is (almost) impossible to use conjugacy so we need to resort to simulative approaches or approximations:
  - Monte Carlo methods
  - Markov Chain Monte Carlo
  - INLA

# Monte Carlo Simulation

- This approach is based on the idea that if you have a large random sample from a certain distribution, the statistics that you can calculate in this sample (mean, SD, percentiles...) will be very similar to the corresponding theoretical values in the distribution.

- If you have a complicated mathematical expression for a distribution and you cannot calculate algebraically important parameters, you could get the computer to generate a large random sample from such a distribution

- By calculating the mean of that parameter in the sample you could estimate the mean in the original distribution with great precision

# Monte Carlo approach to approximate log-odds

- We start with a Binomial likelihood

$$y \mid \theta \sim \text{Binomial}(\theta, n)$$

combined with a

$$\text{Beta}(a, b)$$

as prior for the probability of success $\theta$.

- We are interested in the log-odds function of $\theta$ defined as

$$\log\left(\frac{\theta}{1-\theta}\right)$$

- The integral

$$\int_0^1 \log\left(\frac{\theta}{1-\theta}\right) p(\theta \mid y) \mathrm{d}\theta$$

cannot be computed analytically; we resort to Monte Carlo approximation

# Example of MC: in practice

- We simulate $m$ independent values $\left\{\theta^{(1)}, \ldots, \theta^{(m)}\right\}$ from the

$$\text{Beta}(a_1 = y + a, b_1 = n - y + b)$$

posterior distribution using the property of conjugacy (Beta prior is conjugate to the Binomial likelihood).

- We apply the log-odds transformation to each value obtaining the set of values

$$\left\{\log\left(\frac{\theta^{(1)}}{1-\theta^{(1)}}\right), \ldots, \log\left(\frac{\theta^{(m)}}{1-\theta^{(m)}}\right)\right\}$$

- Finally, we compute the sample mean

$$\frac{\sum_{i=1}^{m} \log\left(\frac{\theta^{(i)}}{1-\theta^{(i)}}\right)}{m}$$

which is the Monte Carlo approximation to

$$\log\left(\frac{\theta}{1-\theta}\right)$$

# Example of MC: R code

In R:

```
> a <- 1
> b <- 1
> theta <- rbeta(1,a,b)
> n <- 1000
> y <- rbinom(1, size=n, p=theta)
> a1 <- a + y
> b1 <- n - y + b
```

# Example of MC: R code

In R:

```
> a <- 1
> b <- 1
> theta <- rbeta(1,a,b)
> n <- 1000
> y <- rbinom(1, size=n, p=theta)
> a1 <- a + y
> b1 <- n - y + b
```

- With this setting the exact posterior distribution of $\theta$ is

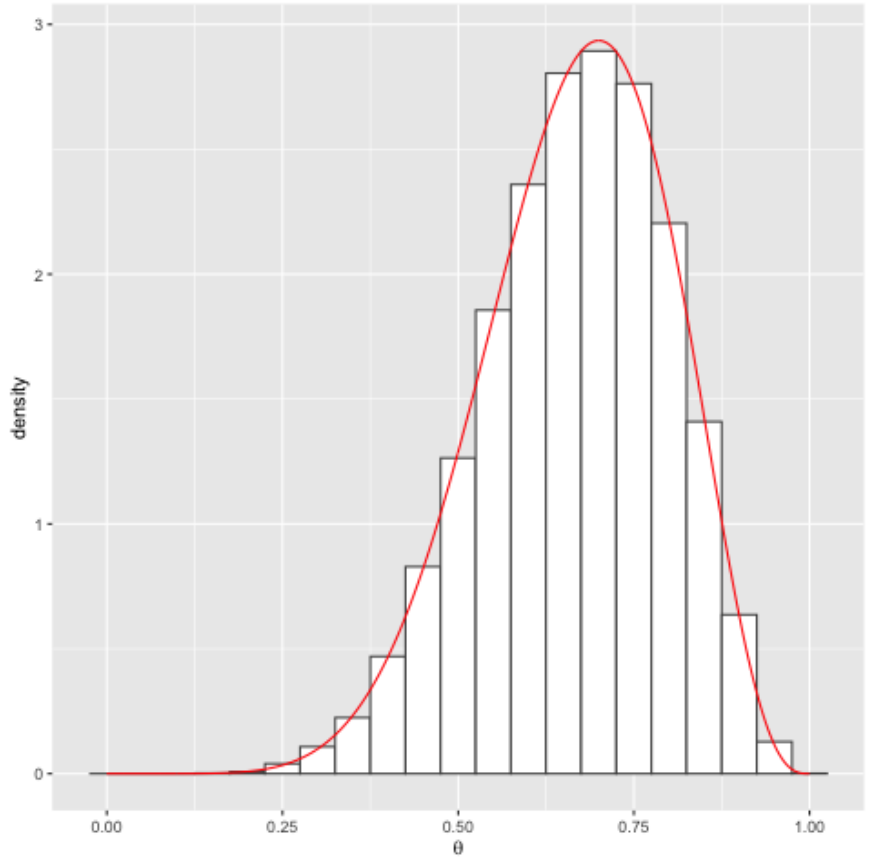
$$\text{Beta}(a_1 = a + y, b_1 = n - y + b)$$

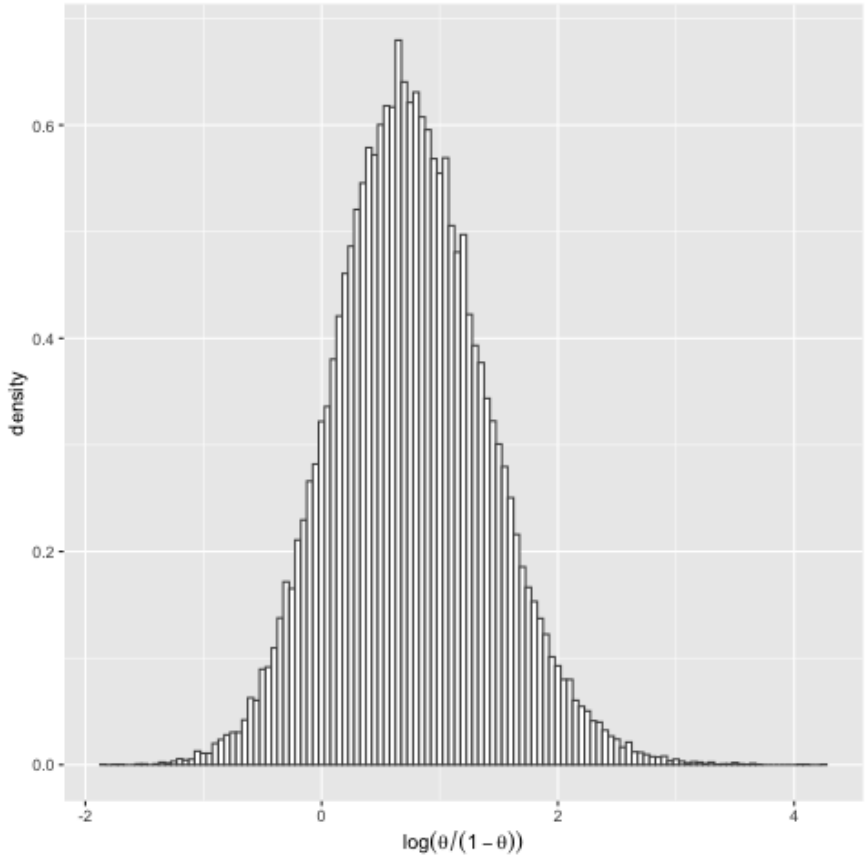
- To approximate the log-odds, we simulate $m = 50000$ values from this Beta posterior distribution using the `rbeta` function.

```
> sim <- rbeta(n=50000, shape1=a1, shape2=b1)
> logodds <- log(sim/(1-sim))
```

# Results and comparison with the theoretical distribution

The empirical distribution of the Monte Carlo sample is plotted below together with the exact posterior distribution of $\theta$.



# Why Markov Chain Monte Carlo?

- For all but trivial examples it will be difficult to draw an iid Monte Carlo sample directly from the posterior distribution.
  – This happens, for example, when the dimension of the parameter vector $\boldsymbol{\theta}$ is high
  – Also to use MC methods we must have a known form for the posterior distribution

# Why Markov Chain Monte Carlo?

- For all but trivial examples it will be difficult to draw an iid Monte Carlo sample directly from the posterior distribution.
  – This happens, for example, when the dimension of the parameter vector $\boldsymbol{\theta}$ is high
  – Also to use MC methods we must have a known form for the posterior distribution

- Alternatively *correlated* values (via MCMC) can be (more easily) drawn to approximate the posterior distribution of the parameter of interest
- Instead of simulating independent values from the posterior distribution we draw a sample by running a Markov chain whose stationary distribution is the posterior density $p(\theta \mid y)$

# Why Markov Chain Monte Carlo?

- For all but trivial examples it will be difficult to draw an iid Monte Carlo sample directly from the posterior distribution.
  – This happens, for example, when the dimension of the parameter vector $\boldsymbol{\theta}$ is high
  – Also to use MC methods we must have a known form for the posterior distribution

- Alternatively *correlated* values (via MCMC) can be (more easily) drawn to approximate the posterior distribution of the parameter of interest
- Instead of simulating independent values from the posterior distribution we draw a sample by running a Markov chain whose stationary distribution is the posterior density $p(\theta \mid y)$

- A sequence of values $\left\{\theta^{(1)}, \ldots, \theta^{(m)}\right\}$ generated from a Markov chain that has reached its stationary distribution (i.e. has converged) can be considered as an approximation to the posterior distribution and can be used to compute all the summaries of interest.

# Why Markov Chain Monte Carlo?

- For all but trivial examples it will be difficult to draw an iid Monte Carlo sample directly from the posterior distribution.
  – This happens, for example, when the dimension of the parameter vector $\boldsymbol{\theta}$ is high
  – Also to use MC methods we must have a known form for the posterior distribution

- Alternatively *correlated* values (via MCMC) can be (more easily) drawn to approximate the posterior distribution of the parameter of interest
- Instead of simulating independent values from the posterior distribution we draw a sample by running a Markov chain whose stationary distribution is the posterior density $p(\theta \mid y)$

- A sequence of values $\left\{\theta^{(1)}, \ldots, \theta^{(m)}\right\}$ generated from a Markov chain that has reached its stationary distribution (i.e. has converged) can be considered as an approximation to the posterior distribution and can be used to compute all the summaries of interest.

- MCMC methods are very general and can effectively be applied to any model
- Even if **in theory**, MCMC can provide (nearly) exact inference, given perfect convergence and MC error $\rightarrow 0$, in practice, this has to be balanced with model complexity and running time
- This is an issue particularly for problems characterised by large data or very complex structure (e.g. hierarchical models)

# MCMC: Gibbs sampling

The **Gibbs sampling** (GS) is one of the most popular schemes for MCMC. Consider the case of a generic $J$ dimensional parameter set $(\theta_1, \theta_2, \ldots, \theta_J)$:

1. Select a set of initial values $(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_J^{(0)})$
2. Sample $\theta_1^{(1)}$ from the conditional distribution $p(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_J^{(0)}, y)$; Sample $\theta_2^{(1)}$ from the conditional distribution $p(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_J^{(0)}, y)$; ...; Sample $\theta_J^{(1)}$ from the conditional distribution $p(\theta_J \mid \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{J-1}^{(1)}, y)$

# MCMC: Gibbs sampling

The **Gibbs sampling** (GS) is one of the most popular schemes for MCMC. Consider the case of a generic $J$ dimensional parameter set $(\theta_1, \theta_2, \ldots, \theta_J)$:

1. Select a set of initial values $(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_J^{(0)})$
2. Sample $\theta_1^{(1)}$ from the conditional distribution $p(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_J^{(0)}, y)$; Sample $\theta_2^{(1)}$ from the conditional distribution $p(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_J^{(0)}, y)$; ...; Sample $\theta_J^{(1)}$ from the conditional distribution $p(\theta_J \mid \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{J-1}^{(1)}, y)$
3. Repeat step 2. for $S$ times until convergence is reached to the target distribution $p(\boldsymbol{\theta} \mid y)$
4. Use the sample from the target distribution to compute all relevant statistics: (posterior) mean, variance, credibility intervals, etc.

If the *full conditionals* are not readily available, they need to be estimated (eg via Metropolis-Hastings) before applying the GS

Easy references for MCMC are

- Blangiardo and Cameletti (2015), Chapter 4
- Johnson, Ott, and Dogucu (2022), Chapters 6-7

# MCMC: convergence

# MCMC: convergence

# MCMC: convergence

# MCMC: pros & cons

- Standard MCMC sampler are generally easy-ish to program and are in fact implemented in readily available software
- MCMC methods are flexible and able to deal with virtually any type of data and model, but they involve computationally- and time- intensive simulations to obtain the posterior distribution for the parameters. For this reason the complexity of the model and the database dimension often remain fundamental issues.

# MCMC: pros & cons

- Standard MCMC sampler are generally easy-ish to program and are in fact implemented in readily available software
- MCMC methods are flexible and able to deal with virtually any type of data and model, but they involve computationally- and time- intensive simulations to obtain the posterior distribution for the parameters. For this reason the complexity of the model and the database dimension often remain fundamental issues.
- The INLA algorithm proposed by (Rue, Martino, and Chopin, 2009) is a *deterministic* algorithm for Bayesian inference and it represents an alternative to MCMC which is instead a simulation based algorithm.

# MCMC: pros & cons

- Standard MCMC sampler are generally easy-ish to program and are in fact implemented in readily available software
- MCMC methods are flexible and able to deal with virtually any type of data and model, but they involve computationally- and time- intensive simulations to obtain the posterior distribution for the parameters. For this reason the complexity of the model and the database dimension often remain fundamental issues.
- The INLA algorithm proposed by (Rue, Martino, and Chopin, 2009) is a *deterministic* algorithm for Bayesian inference and it represents an alternative to MCMC which is instead a simulation based algorithm.
- The INLA algorithm is designed for the class of *latent Gaussian models* and compared to MCMC it provides (as) accurate results in a shorter time.

# MCMC: pros & cons

- Standard MCMC sampler are generally easy-ish to program and are in fact implemented in readily available software
- MCMC methods are flexible and able to deal with virtually any type of data and model, but they involve computationally- and time- intensive simulations to obtain the posterior distribution for the parameters. For this reason the complexity of the model and the database dimension often remain fundamental issues.
- The INLA algorithm proposed by (Rue, Martino, and Chopin, 2009) is a *deterministic* algorithm for Bayesian inference and it represents an alternative to MCMC which is instead a simulation based algorithm.
- The INLA algorithm is designed for the class of *latent Gaussian models* and compared to MCMC it provides (as) accurate results in a shorter time.
- INLA has become very popular amongst statisticians and applied researchers and in the past few years the number of papers reporting usage and extensions of the INLA method has increased considerably.

# References

Blangiardo, M. and M. Cameletti (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Johnson, A. A., M. Q. Ott, and M. Dogucu (2022). *Bayes Rules!: An Introduction to Applied Bayesian Modeling*. CRC Press.

Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian inference for latent Gaussian model by using integrated nested Laplace approximations (with discussion)". In: *J. R. Statist. Soc. B* 71, pp. 319-392.