UPPSALA
UNIVERSITET

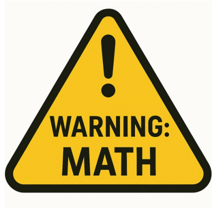
Introduction to Bayesian modeling

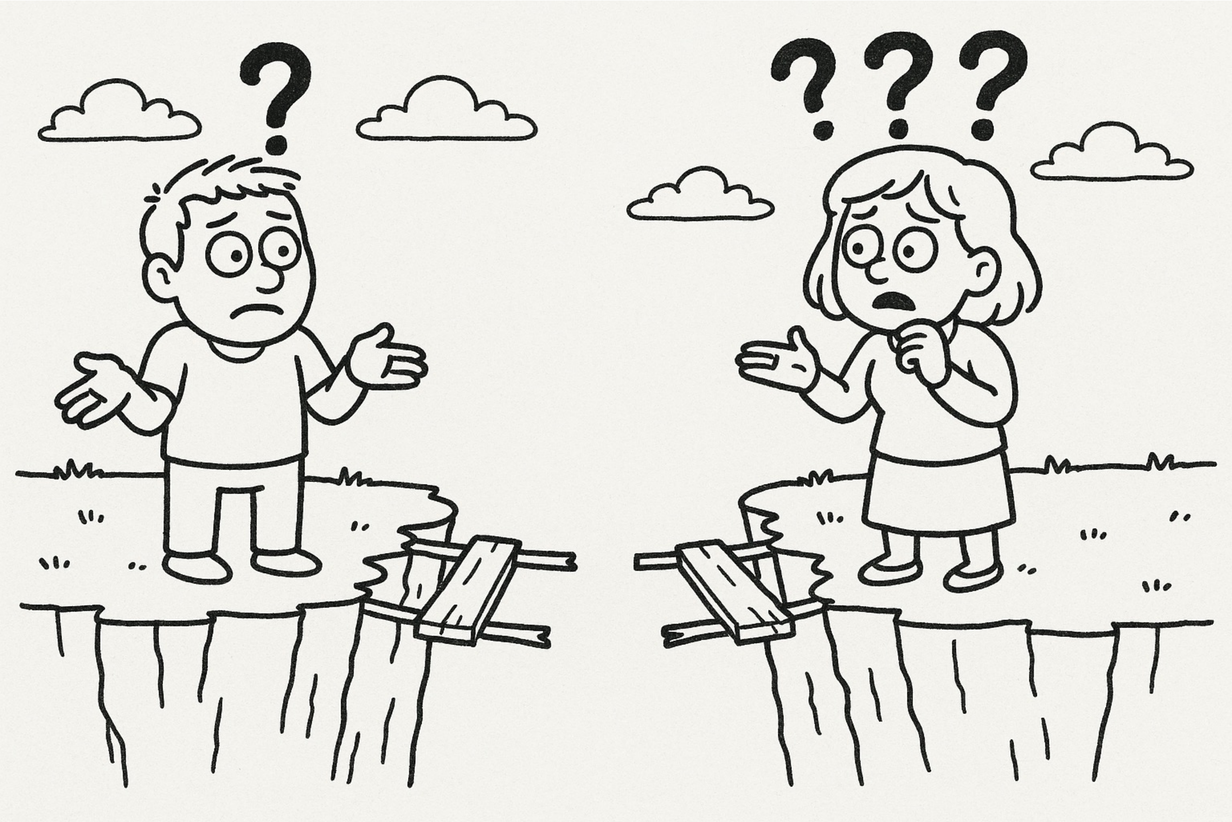
Jakob Torgander

Outline:

- Introduction to Bayesian statistics
- Model construction & comparison
- Bayesian computation
- Demo

Theory to practice: Bayesian bridge





UPPSALA
UNIVERSITET

About me

- PhD Student in Statistics
- Research on Bayesian statistics & Probabilistic programming
- Part of the Interdisiplinary Bayes (InterBayes) network at Uppsala.
- Co-developer of *posteriordb:* a framework for testing and benchmarking Bayesian algorithms

UPPSALA
UNIVERSITET

Introduction to Bayesian Statistcs

UPPSALA
UNIVERSITET

Why Bayesian statistics?

- Flexible framework for statistical modeling
- Incorporates prior information → Less data needed
- Builds on explicit model assumptions rather than implicit
- Framework for updating beleifs as new information is collected
- Directly models uncertainty of parameters and predictions → Intuitive inference

Why Bayesian statistics?

Which of the following is a correct charaterization of a (1-α) % confidence interval [a,b]?

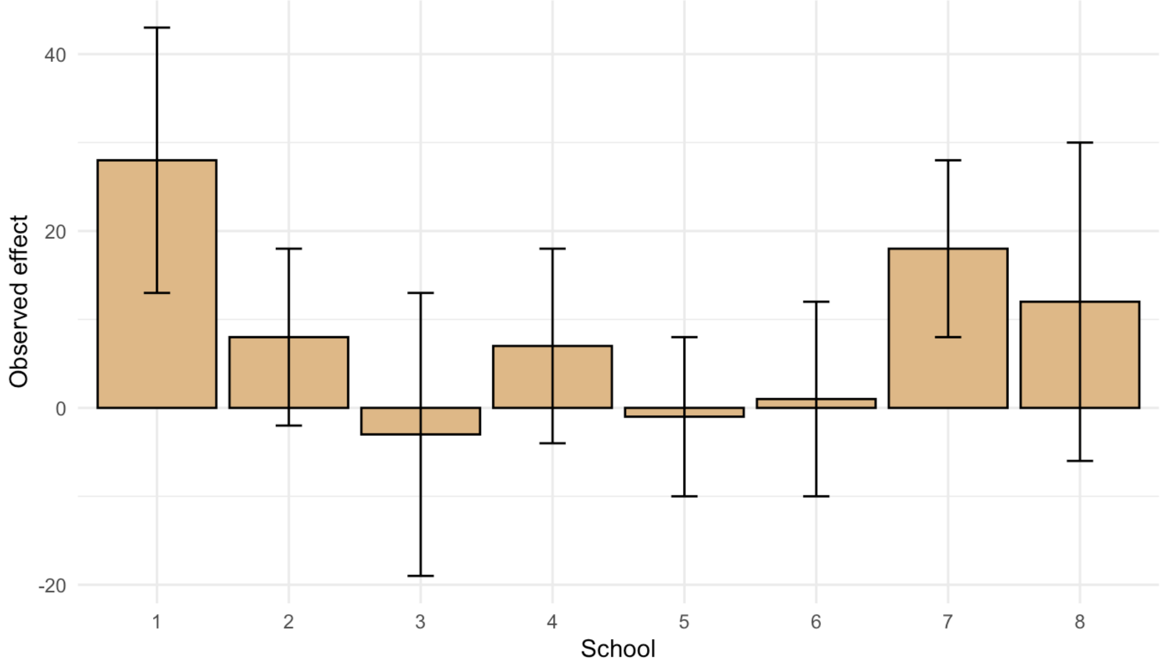
a) The probability that the null hypothesis is true is (1-α) %
b) If I repeat my experiment infiniteley many times, the resulting intervals will contain the true parameter (1-α) % of the time
c) The probability that the true parameter is between a and b is (1-α) %

Bayesian Credible intervals!

UPPSALA UNIVERSITET

Case study – “Eight schools” data

- Studying the effect of coaching programs on SAT-V scores in eight schools
- Data: Treatment effect and standard errors for each school
- Research questions: Global vs local effects?
- Statistical problem: few observations, high uncertainty



UPPSALA
UNIVERSITET

Posterior distribution

- Main object for Bayesian statistics: posterior distribution of parameter θ given data x
- Posterior is defined mathematically as:

$$\mathrm{p}(\theta|\mathrm{x}) = \frac{p(x|\theta)p(\theta)}{Z}$$

Bayes' rule ($\mathrm{Z} = \mathrm{p}(\mathrm{x})$)

- Has three components:
  - $p(x|\theta)$: Likelihood – relationship between data and parameter
  - $p(\theta)$: Prior – Apriori knowledge of θ (if any)
  - Z = p(x): Constant only depending on data. We'll worry about this later.
- Choosing and computing these components is the focus of <u>Bayesian</u> modeling!

UPPSALA UNIVERSITET

Eight schools example: Baseline model

- Model assumption: Given the true treatment θ_i for school i the observed change in SAT score y_i is normally distributed

$$y_i | \theta_i, {\sigma_i}^2 \sim \mathcal{N}(\theta_i, {\sigma_i}^2)$$

- A simple (weakly informative) prior for $\theta_{:i}$

$$\theta_i \sim \mathcal{N}(0, 5^2)$$

- Can be shown that the resulting posterior is also normal:

$$\theta_i | y_i \sim \mathcal{N}(\theta_{post}, \sigma_{post}^2),$$

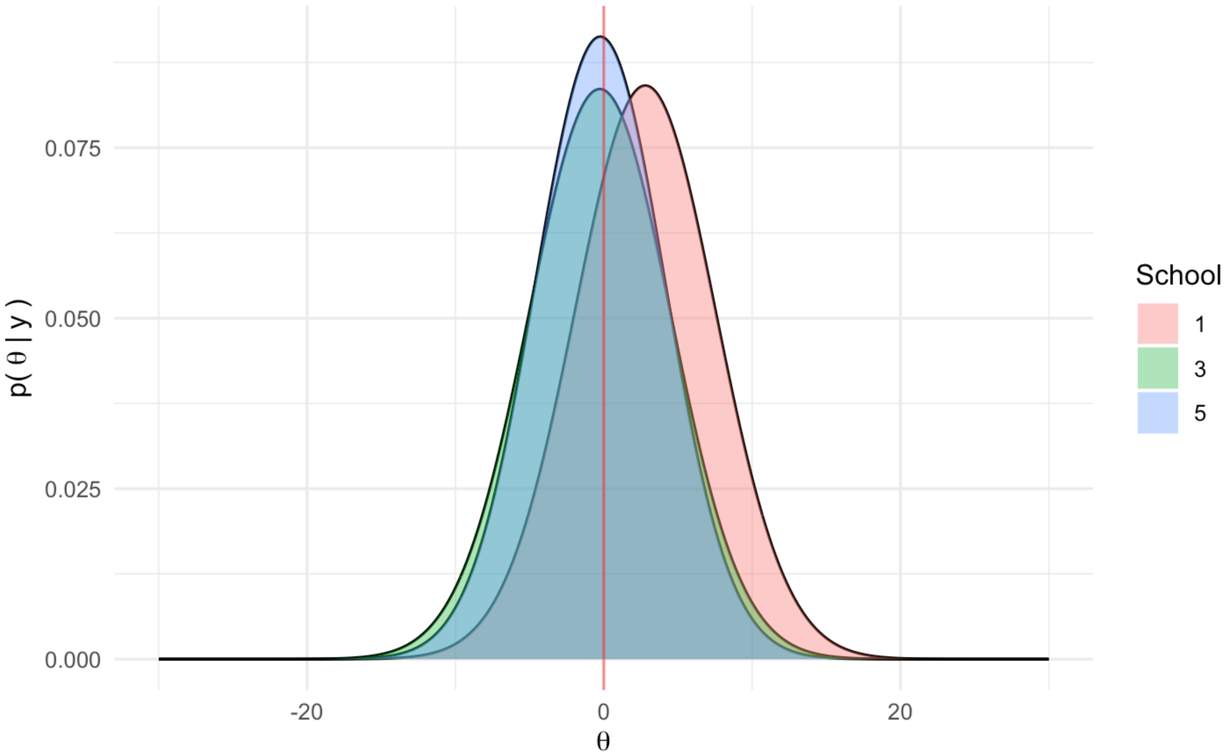
where

$$\theta_{post} = \frac{y_i}{\sigma_i^2 (\frac{1}{25} + \frac{1}{\sigma_i^2})}, \quad \sigma_{post}^2 = \left(\frac{1}{25} + \frac{1}{\sigma_i^2} \right)^{-1}$$

Note: In general, posterior distributions are not easy to derive analytically!

UPPSALA UNIVERSITET

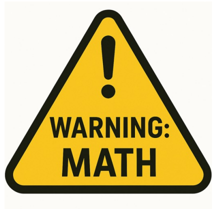
Eight schools example: Baseline model



- All posteriors overlapping zero: No “significant” effect of SAT-coaching using this model
- Uncertainty very high due to few observations!
- Will later see how “sharing” information between parameters gives a better model

UPPSALA UNIVERSITET

Why posterior distribution?



- Key concept: **Posterior contains all information needed for inference/prediction**
- Mathematically: compute expected values:

$$\mathbb{E}[h(\theta)] = \int h(\theta) p(\theta|x) d\theta$$

  for some “informative” function h

- Most common choises for h:
  - $h(\theta) = \theta$: **Posterior mean -** “Expected parameter value”
  - $h(\theta) = 1(a < \theta < b)$: **Credible interval -** “Probability that θ is between a and b”
  - $h(\theta) = p(\hat{y}|\theta)$ predictive distribution for new data point

UPPSALA
UNIVERSITET

Predictive distrbution

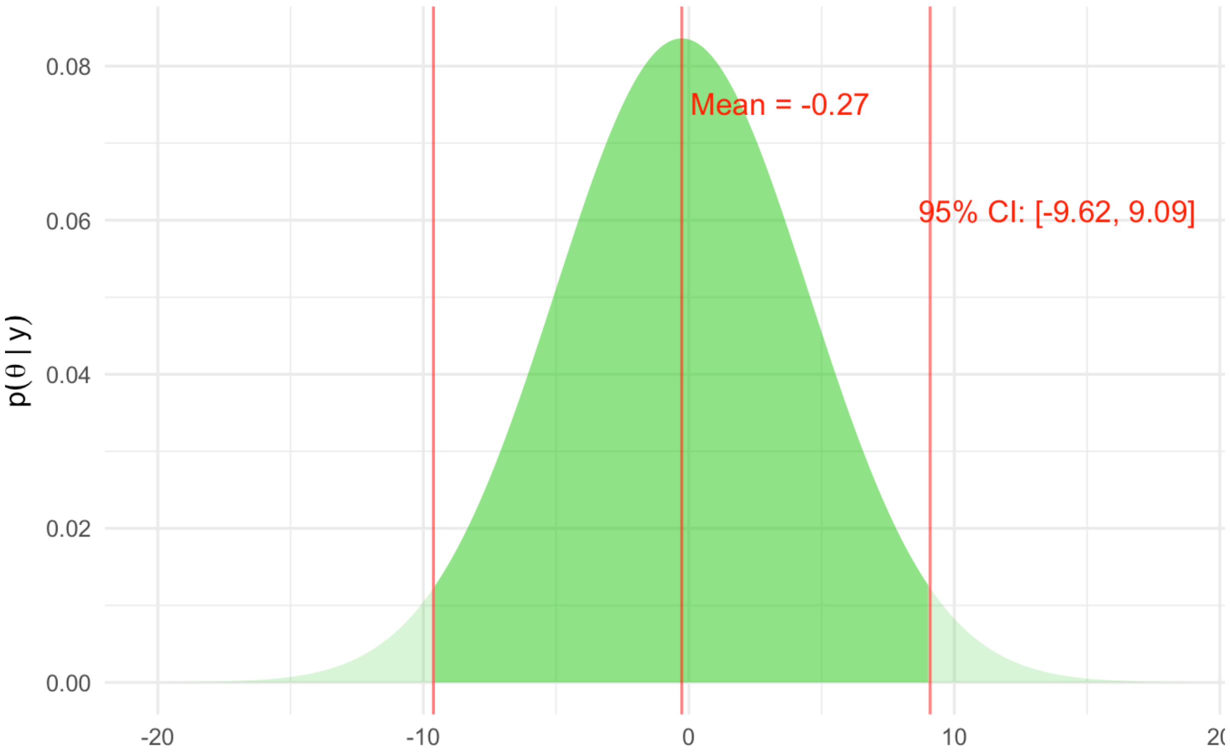
- The posterior can also be used for generating predictions and predictions intervals
- Letting $h(\theta) = p(\hat{y}|\theta)$ we get the predictive density of a new point $\hat{y}$ by the following

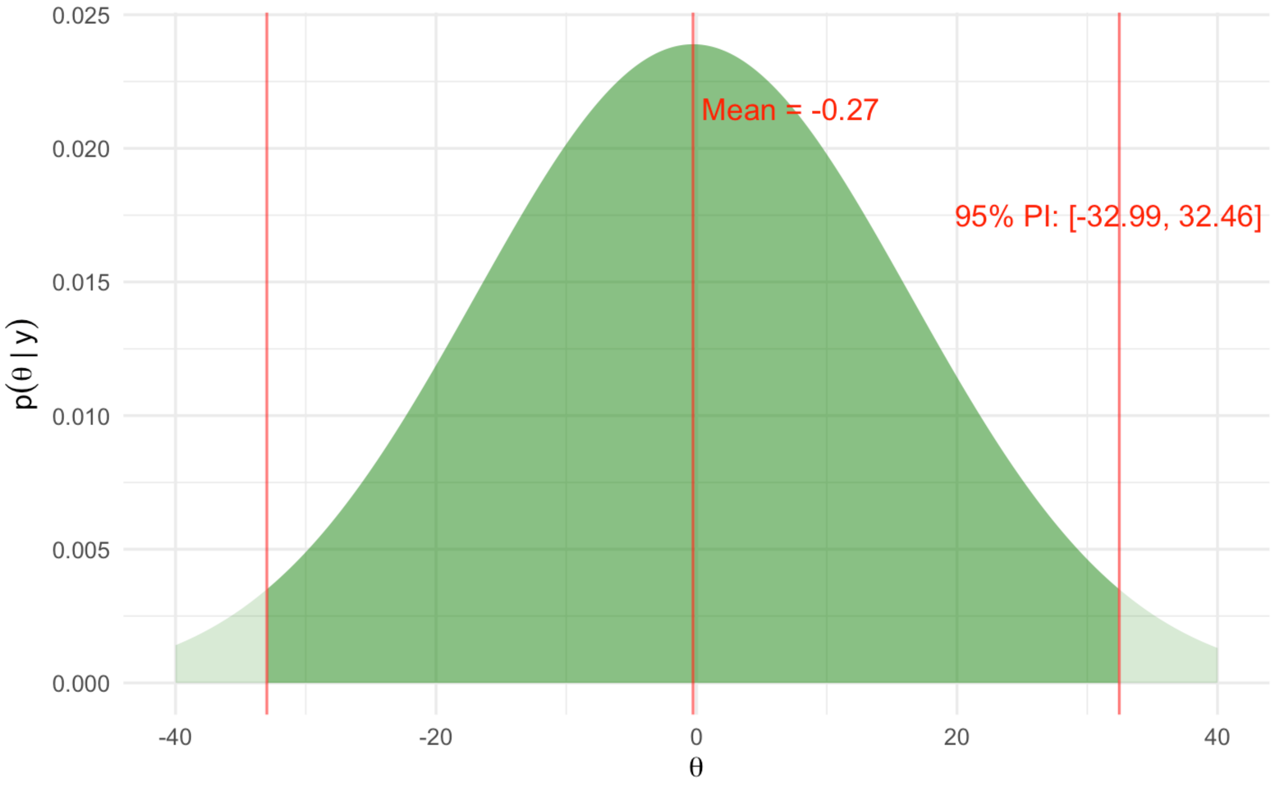
$$p_{post}(\hat{y}|\theta) = \int p(\hat{y}|\theta)p(\theta|y)dy$$

- Will see how this quantity can be used for comparing models

UPPSALA
UNIVERSITET

School # 3: 95% Credible and prediction intervals





UPPSALA
UNIVERSITET

Model construction & comparison

UPPSALA
UNIVERSITET

Model construction

- Recall: to define a posterior

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{Z},$$

we need to define a likelihood and prior component.

- Often a combination of standard probability distributions
- Main approaches:
  - Previous work/litterature
  - Historically: Restrict to “easy” posteriors
  - Domain expertiese, experimentation and model comparison
- Good starting point: begin with standard model for problem and add weakly informative (high variance) prior

Example: Linear regression

- Consider the standard univariate linear regression model

$$Y_i = \alpha + \beta X_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- A corresponding Bayesian model with weakly informative priors can be defined as follows:

$$\begin{gathered} Y_i | X_i, \alpha, \beta \sim \mathcal{N}(\alpha + \beta X_i, \sigma^2) \\ \alpha \sim \mathcal{N}(0, 10) \\ \beta \sim \mathcal{N}(0, 10) \\ \sigma \sim \mathcal{N}^+(0, 10) \end{gathered}$$

UPPSALA
UNIVERSITET

Eight schools: Hierarchical model

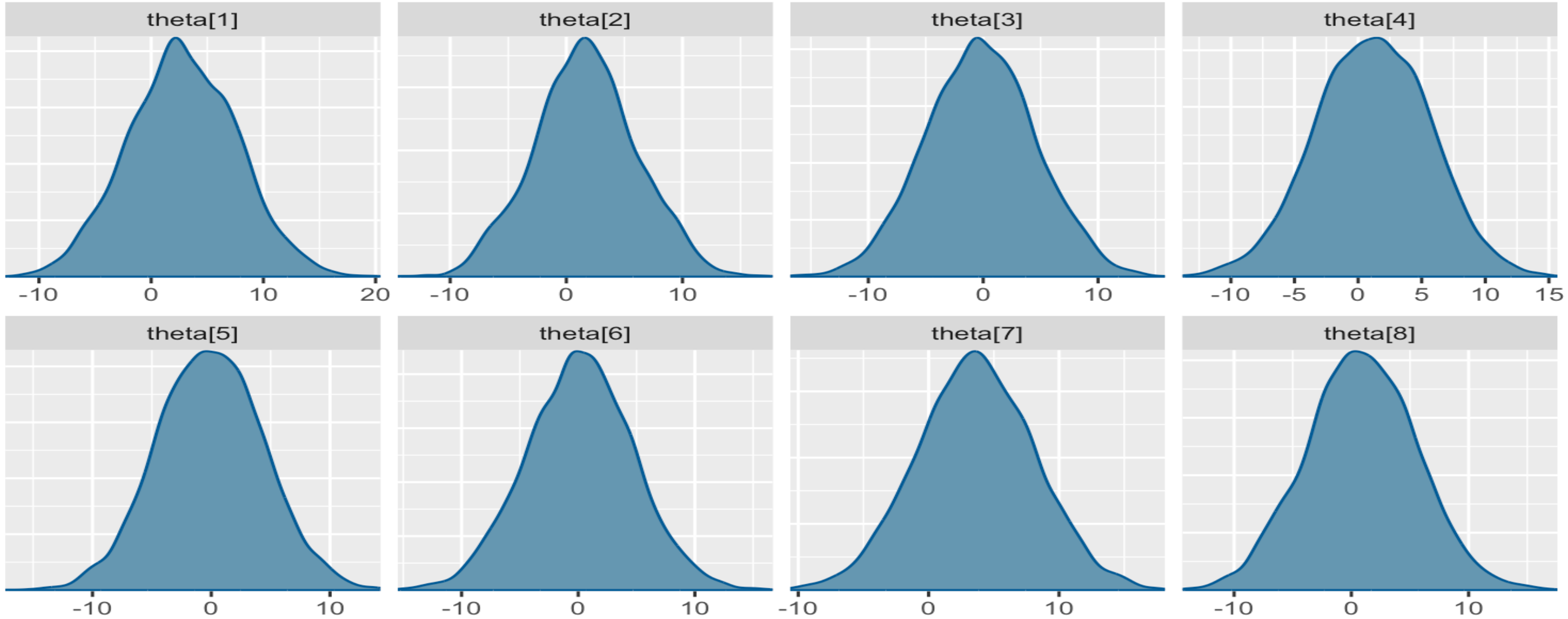
- Model construction is often done iteratively using information from a previous model.
- Recall: Few observations results in high uncertainty for baseline eight schools model
- Idea for better model: share information between parameters by putting ”prior on prior”
- Results in a **hierarchical** model:

$$y_i|\theta_i, {\sigma_i}^2 \sim \mathcal{N}(\theta_i, {\sigma_i}^2)$$

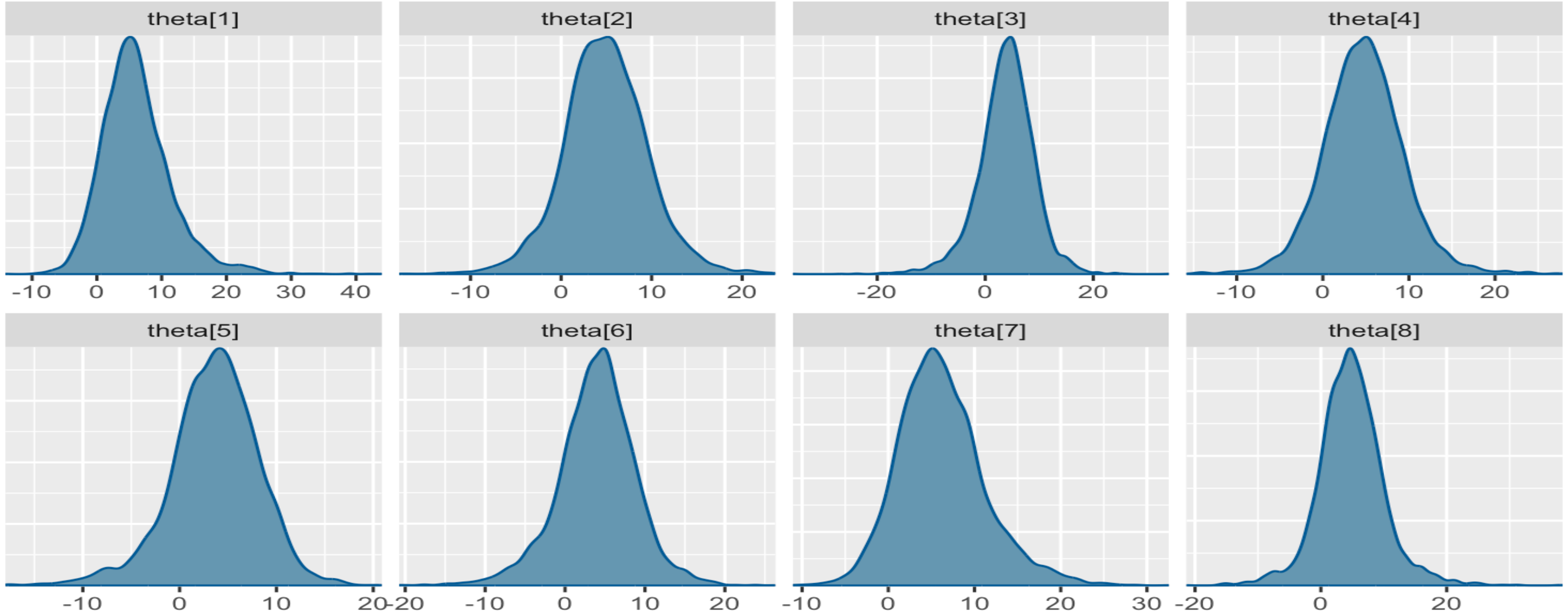
$$\theta_i \sim \mathcal{N}(\mu, \tau^2)$$

$$\mu \sim \mathcal{N}(0, 5^2)$$

Baseline:



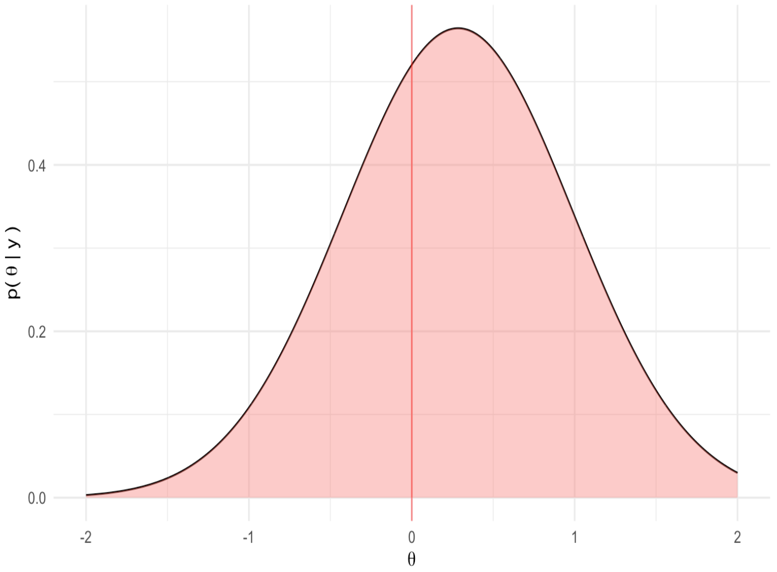
Hierarchical:



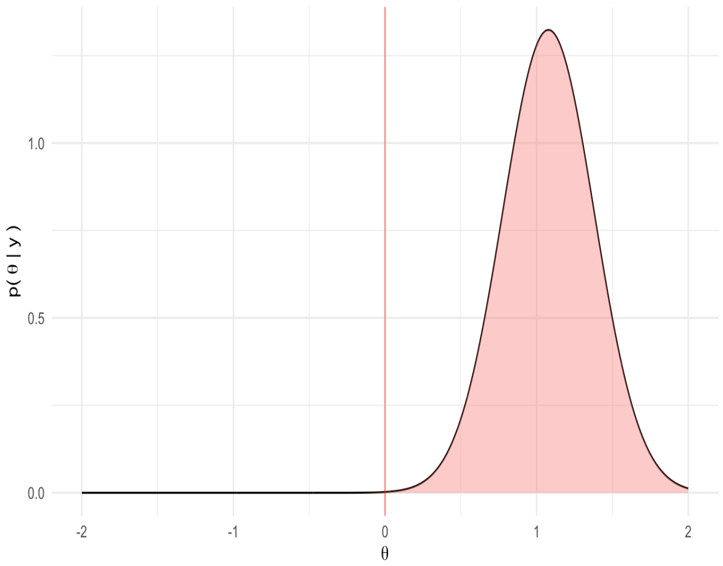
UPPSALA UNIVERSITET

Prior distributions

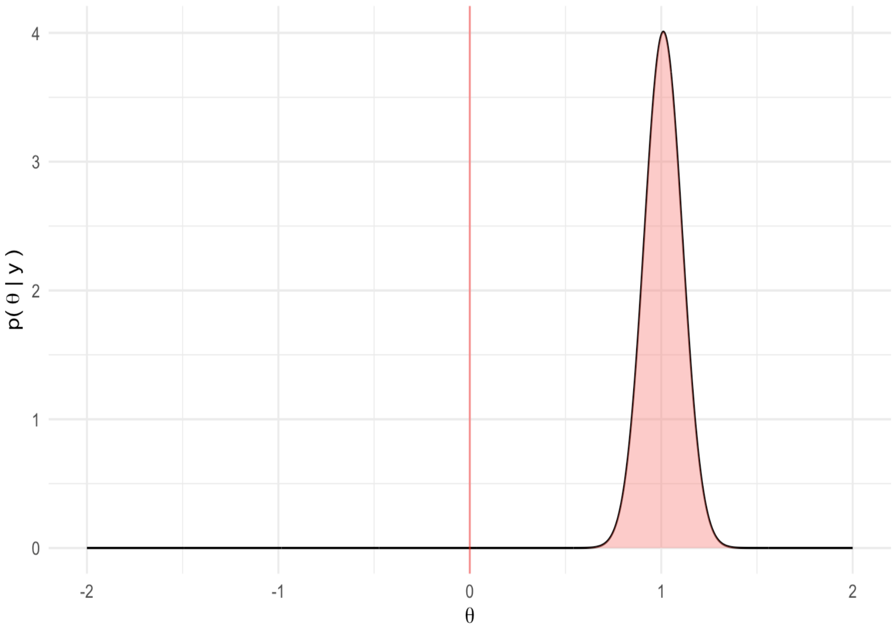
- Should/could reflect previous knowledge about the parameter
- Can be chosen to be informative or (weakly) informative
- Important concept: if we have enough data the likelihood will dominate the prior and vice versa. Illustration for the normal-normal posterior:



n = 1



n = 10



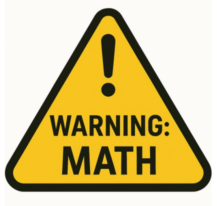
n = 100

Common likelihood/prior combinations

Data	Parameter	Likelihood	Prior
Continuous	Mean Variance	Normal	Normal Inv-Gamma
Discrete	Mean Variance	Poisson Neg.binomial	Normal Normal
Binary	Prob. of success Treatment effect	Bernoulli Logit	Beta Dirichlet Normal

UPPSALA
UNIVERSITET

Model comparison



- Conceptually: Compare models based on how well they predict future data
- Recall: the predicitve density:

$$p_{post}(\hat{y}|\theta) = \int p(\hat{y}|\theta)p(\theta|y)dy$$

- Idea: compare candidate models using the Expected Log Predictive Denisty (ELPD):

$$\mathbb{E}[\log(p_{post}(\hat{Y}|\theta))] = \int \log(p_{post}(\hat{y}|\theta))p(\hat{y})d\hat{y}$$

- Compares how close the predictive denisty is to the true (unknown) data generating process $p(\hat{y})$
- Can be estimated efficiently with cross validation techniques using software

Bayesian computation

UPPSALA
UNIVERSITET

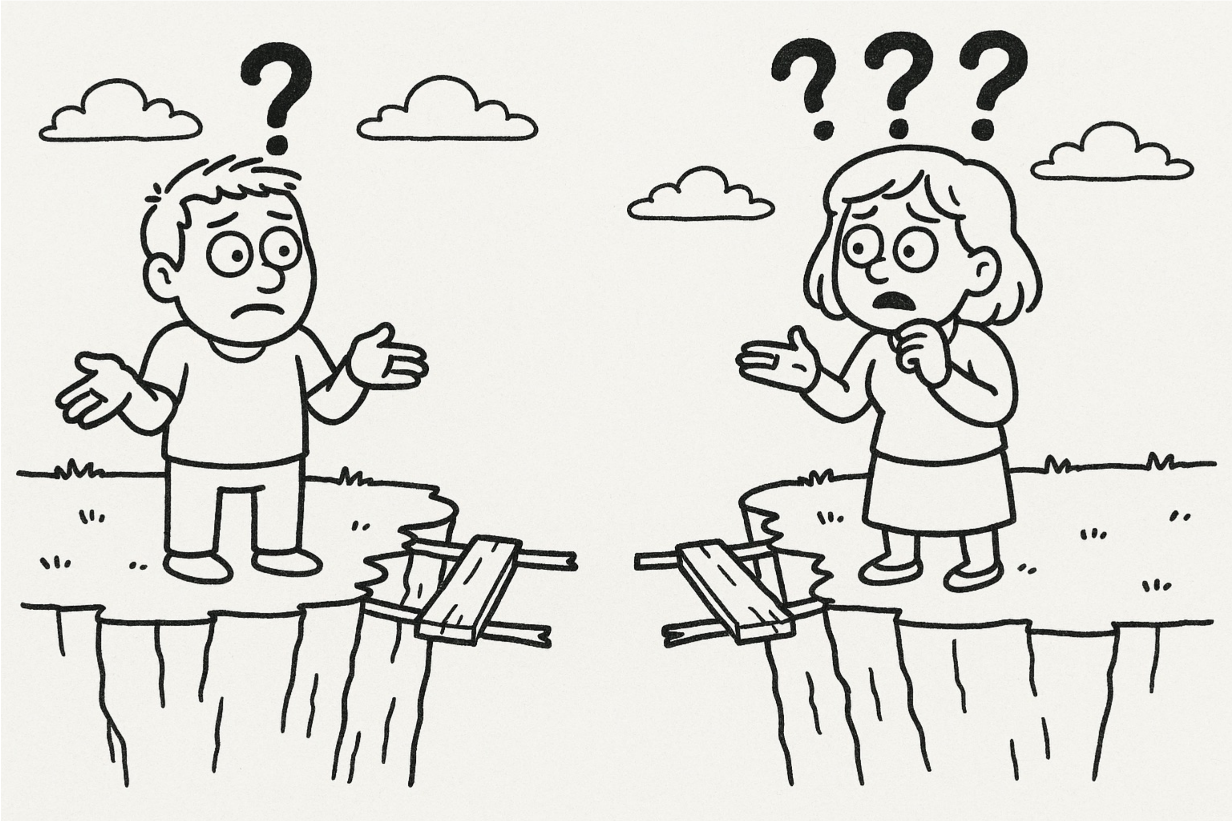
Next step: Bayesian computation

- Recall: posterior definition $\mathrm{p}(\theta|\mathrm{x}) = \frac{p(x|\theta)p(\theta)}{Z}$ depends on constant

$$\mathrm{Z} = \mathrm{p}(\mathrm{x}) = \int \mathrm{p}(\mathrm{x} \mid \theta)\mathrm{p}(\theta)\, \mathrm{d}\theta$$

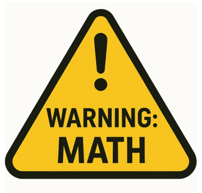
- Problem: Z is in general intractable → Closed form of $\mathrm{p}(\theta|\mathrm{x})$ rarely exists.

- Circumventing this problem is the main objective of Bayesian computation and Probabllisitic Programming Languages (e.g. Stan, Pyro, PyMC,Turing.jl, etc).



UPPSALA
UNIVERSITET

Monte Carlo methods:



- Idea: instead of computing full posterior (hard), focus on the expectations (easier)

$$\mathbb{E}[h(\theta)] = \int h(\theta)p(\theta|x)d\theta$$

- Monte Carlo integration: estimate using simulated draws $\theta^{(1)}, \ldots, \theta^{(m)}$ from the posterior:

$$\mathbb{E}[h(\theta)] \approx \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)})$$

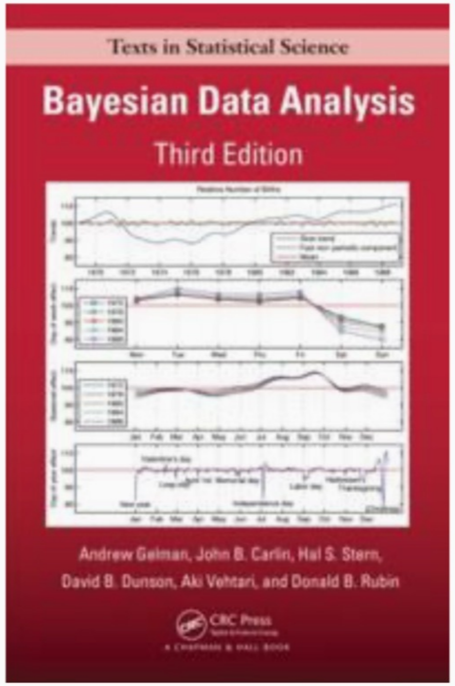
- For drawing the samples we use Stan

Stan

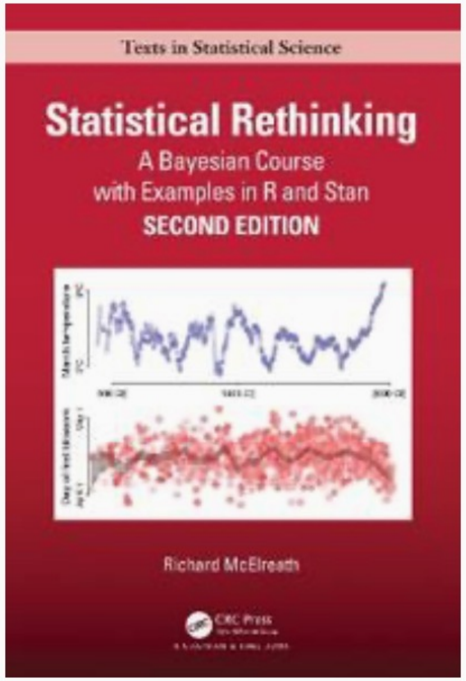
- Released in 2012
- Currently most popular language for Bayesian computation
- Link to documentation & tutorials: https://mc-stan.org/docs/
- Special Interbayes/CIRCUS Workshop in Stan 6/11

Stan demo

Resources:



Bayesian Data Analysis, 3rd ed
Gelman et. al (2020)
CRC Press



Statistical Rethinking, 2nd ed
Richard McElreath (2018)
CRC Press

UPPSALA
UNIVERSITET

InterBayes – Upcoming events

- Workshop: Introduction to Stan – 6/11
- Stan Hackaton 10/11
- StanCon 2026
- Website: https://interbayes.github.io
- Link to slides and code: https://github.com/JTorgander/interbayes-workshop/

UPPSALA
UNIVERSITET