

# Bayesian Learning

## Lecture 1 - The Bayesics and Bernoulli data

**Mattias Villani**

Department of Statistics  
Stockholm University

Department of Computer and Information Science  
Linköping University



[mattiasvillani.com](http://mattiasvillani.com)



@matvil



mattiasvillani

# Course overview

■ Course [webpage](#). Course [syllabus](#).

■ Modes of teaching:

- ▶ Lectures ([Mattias Villani](#))
- ▶ Mathematical exercises ([Oscar Oelrich](#))
- ▶ Computer labs (Oscar Oelrich)

■ **Modules:**

- ▶ The **Bayesics**, single- and multiparameter models
- ▶ **Regression** and **Classification models**
- ▶ **Advanced models** and **Posterior Approximation** methods
- ▶ **Model Inference** and **Variable Selection**

■ **Examination**

- ▶ Lab reports
- ▶ Home exam

# Lecture overview

- The likelihood function
- Bayesian inference
- Bernoulli model

# Likelihood function - Bernoulli trials

## ■ Bernoulli trials:

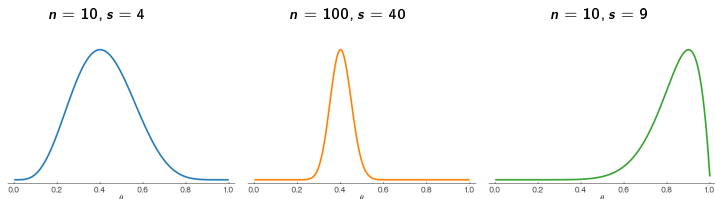
$$X_1, \dots, X_n | \theta \overset{iid}{\sim} \text{Bern}(\theta).$$

## ■ Likelihood from $s = \sum_{i=1}^n x_i$ successes and $f = n - s$ failures.

$$p(x_1, \dots, x_n | \theta) = p(x_1 | \theta) \cdots p(x_n | \theta) = \theta^s (1 - \theta)^f$$

## ■ Maximum likelihood estimator $\hat{\theta}$ maximizes $p(x_1, \dots, x_n | \theta)$ .

## ■ Given the data $x_1, \dots, x_n$ , plot $p(x_1, \dots, x_n | \theta)$ as a function of $\theta$ .



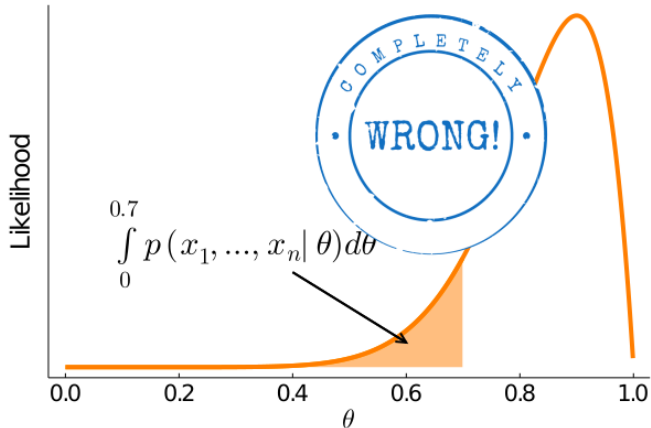
# The likelihood function

- Say it out loud:

*The likelihood function is  
the probability of the observed data  
considered as a function of the parameter.*

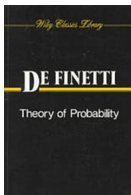
- The symbol  $p(x_1, \dots, x_n | \theta)$  plays two different roles:
- **Probability distribution** for the data.
  - ▶ The data  $x = (x_1, \dots, x_n)$  are random.
  - ▶  $\theta$  is fixed.
- **Likelihood function** for the parameter
  - ▶ The data  $x = (x_1, \dots, x_n)$  are fixed.
  - ▶  $p(x_1, \dots, x_n | \theta)$  is function of  $\theta$ .

# Probabilities from the likelihood?



# Uncertainty and subjective probability

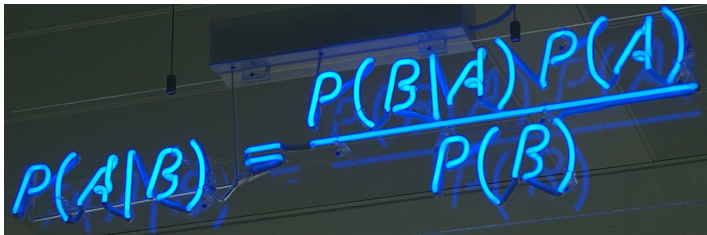
- $\Pr(\theta < 0.6 | \text{data})$  only makes sense if  $\theta$  is random.
- But  $\theta$  may be a fixed natural constant?
- **Bayesian: doesn't matter if  $\theta$  is fixed or random.**
- Do **You** know the value of  $\theta$  or not?
- $p(\theta)$  reflects Your knowledge/**uncertainty** about  $\theta$ .
- **Subjective probability.**
- The statement  $\Pr(10\text{th decimal of } \pi = 9) = 0.1$  makes sense.



# Bayesian learning

- **Bayesian learning** about a model parameter  $\theta$ :
  - ▶ state your **prior** knowledge as a probability distribution  $p(\theta)$ .
  - ▶ collect **data**  $x$  and form the **likelihood** function  $p(x|\theta)$ .
  - ▶ **combine** prior knowledge  $p(\theta)$  with data information  $p(x|\theta)$ .
- **How to combine** the two sources of information?

## Bayes' theorem



A photograph of a chalkboard with the equation for Bayes' theorem written in blue chalk. The equation is 
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
 The chalkboard has a dark background, and the blue chalk is clearly visible. There are some faint, illegible markings on the board, possibly from previous lessons.



# Learning from data - Bayes' theorem

- How to **update** from **prior**  $p(\theta)$  to **posterior**  $p(\theta|Data)$ ?
- **Bayes' theorem** for events  $A$  and  $B$

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

- Bayes' Theorem for a model parameter  $\theta$

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)}.$$

- It is the prior  $p(\theta)$  that takes us from  $p(Data|\theta)$  to  $p(\theta|Data)$ .
- A probability distribution for  $\theta$  is extremely useful.  
**Predictions. Decision making.**
- **No prior - no posterior - no useful inferences - no fun.**

# Medical diagnosis

- $A = \{\text{Very rare disease}\}$ ,  $B = \{\text{Positive medical test}\}$ .
- $p(A) = 0.0001$ .  $p(B|A) = 0.9$ .  $p(B|A^c) = 0.05$ .
- Probability of being sick when test is positive:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A^c)p(A^c)} \approx 0.0018.$$

- Probably not sick, but 18 times more probable now.
- **Morale:** If you want  $p(A|B)$  then  $p(B|A)$  does not tell the whole story. The prior probability  $p(A)$  is also very important.

***“You can’t enjoy the Bayesian omelette  
without breaking the Bayesian eggs”***

*Leonard Jimmie Savage*



# The normalizing constant is not important

- Bayes theorem

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)} = \frac{p(Data|\theta)p(\theta)}{\int_{\theta} p(Data|\theta)p(\theta)d\theta}.$$

- Integral  $p(Data) = \int_{\theta} p(Data|\theta)p(\theta)d\theta$  can make you cry.

- $p(Data)$  is **only a constant** so that  $\int p(\theta|Data) = 1$ .

- Example:  $x \sim N(\mu, \sigma^2)$

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp \left[ -\frac{1}{2\sigma^2} (x - \mu)^2 \right].$$

- We may write

$$p(x) \propto \exp \left[ -\frac{1}{2\sigma^2} (x - \mu)^2 \right].$$

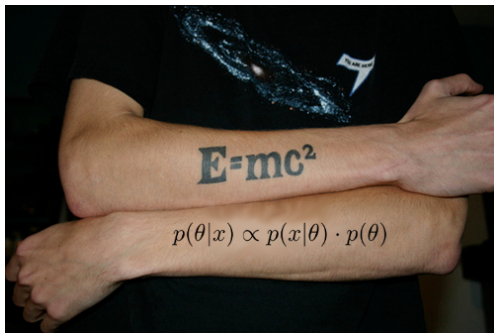
# Great theorems make great tattoos

- All you need to know:

$$p(\theta|Data) \propto p(Data|\theta)p(\theta)$$

or

$$\text{Posterior} \propto \text{Likelihood} \cdot \text{Prior}$$



# Bernoulli trials - Beta prior

## ■ Model

$$x_1, \dots, x_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta)$$

## ■ Prior

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \quad \text{for } 0 \leq \theta \leq 1.$$

## ■ Posterior

$$\begin{aligned} p(\theta | x_1, \dots, x_n) &\propto p(x_1, \dots, x_n | \theta) p(\theta) \\ &\propto \theta^s (1 - \theta)^f \theta^{\alpha-1} (1 - \theta)^{\beta-1} \\ &= \theta^{s+\alpha-1} (1 - \theta)^{f+\beta-1}. \end{aligned}$$

■ Posterior is proportional to the  $\text{Beta}(\alpha + s, \beta + f)$  density.

■ The prior-to-posterior mapping:

$$\theta \sim \text{Beta}(\alpha, \beta) \xrightarrow{x_1, \dots, x_n} \theta | x_1, \dots, x_n \sim \text{Beta}(\alpha + s, \beta + f)$$

# Beta distribution

$X \sim \text{Beta}(\alpha, \beta)$  for  $X \in [0, 1]$ .

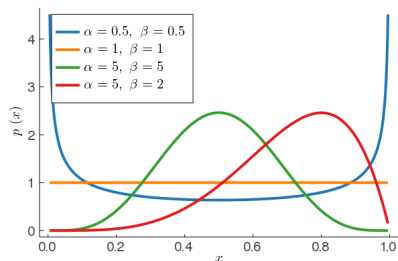
$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta}$$

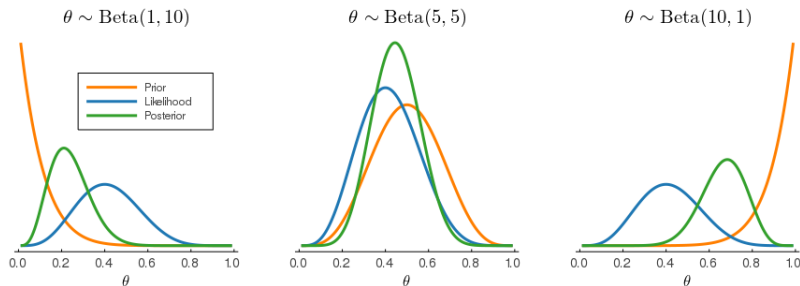
$$\mathbb{V}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

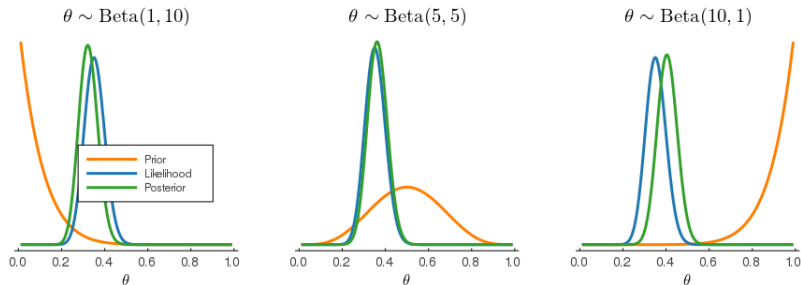
$\Gamma(\alpha)$  is the Gamma function.



# Spam data (n=10) - Prior is influential

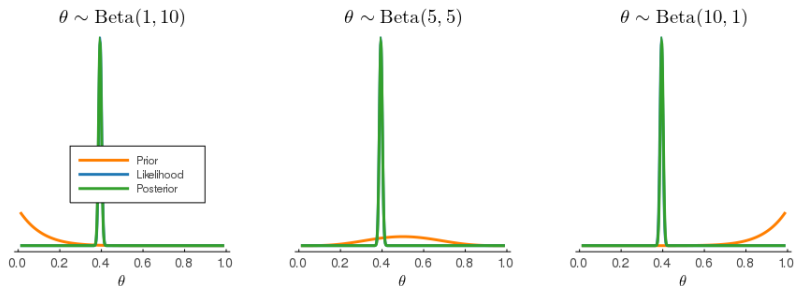


# Spam data ( $n=100$ ) - Prior is less influential





# Spam data (n=4601) - Prior does not matter



# Bayes respects the Likelihood Principle

## ■ Bernoulli trials with order:

$$x_1 = 1, x_2 = 0, \dots, x_4 = 1, \dots, x_n = 1$$

$$p(x|\theta) = \theta^s(1-\theta)^f$$

## ■ Bernoulli trials without order. $n$ fixed, $s$ random.

$$p(s|\theta) = \binom{n}{s} \theta^s (1-\theta)^f$$

## ■ Negative binomial sampling: sample until you get $s$ successes. $s$ fixed, $n$ random.

$$p(n|\theta) = \binom{n-1}{s-1} \theta^s (1-\theta)^f$$

## ■ The **posterior distribution is the same** in all three cases.

## ■ Bayesian inference respects the **likelihood principle**.