# Bayesian Learning

## Lecture 1 - The Bayesics, Bernoulli and Normal data

**Mattias Villani**

Department of Statistics
Stockholm University

mattiasvillani.com    @matvil    @matvil    mattiasvillani

# Course overview

- Course webpage. Course syllabus.
- Modes of teaching:
  - Lectures (Mattias Villani)
  - Mathematical exercises (Oskar Gustafsson)
  - Computer labs to work on Home Assignment (Oskar Gustafsson, Akram Mahmoudi and Valentin Zulj)
- **High-level contents**:
  - The **Bayesics**, single- and multiparameter models
  - **Regression** and **Classification models**
  - **Advanced models** and **Posterior Approximation**
  - **Simulation-based inference** and **Probabilistic programming**
- **Examination**
  - **Home assignment**, Part A and B.
  - **Exam**: Pen and paper + Computer (using R).
    *Polished home assignment can be uploaded as pdf.*

# Lecture overview

- The **likelihood function**
- **Bayesian inference**
- **Bernoulli model**
- The **Normal model** with known variance

# Likelihood function - Bernoulli trials

- **Bernoulli trials**:

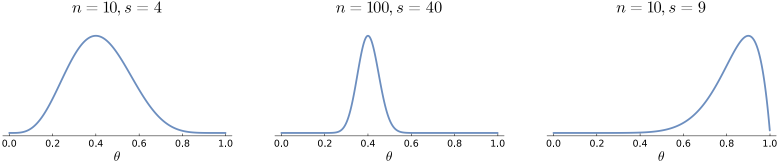
$$X_1, ..., X_n | \theta \overset{iid}{\sim} Bern(\theta).$$

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

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

- **Maximum likelihood estimator** $\hat{\theta}$ maximizes $p(x_1, ..., x_n|\theta)$.

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



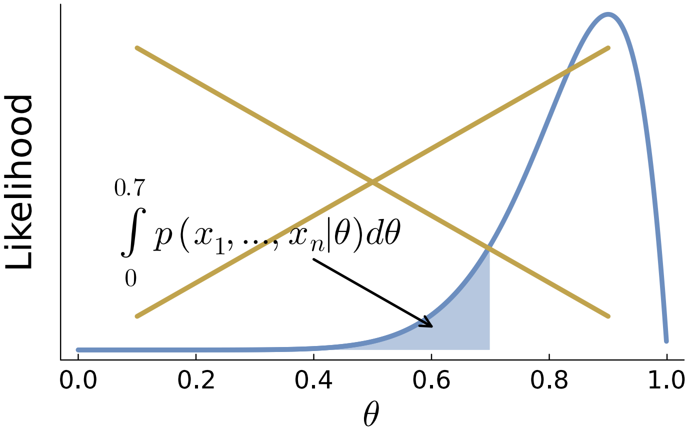
# The likelihood function

- Say it out loud:

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

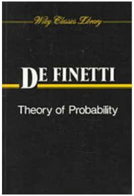
- The symbol $p(x_1, ..., x_n|\theta)$ plays two different roles:
- **Probability distribution** for the data.
  - The data $\mathbf{x} = (x_1, ..., x_n)$ are random.
  - $\theta$ is fixed.
- **Likelihood function** for the parameter
  - The data $\mathbf{x} = (x_1, ..., x_n)$ are fixed.
  - $p(x_1,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

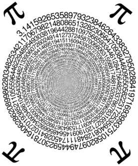
# Probabilities from the likelihood?



# 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** and form the **likelihood** function $p(\mathrm{Data}|\theta)$.
  - **combine** prior $p(\theta)$ and data information $p(\mathrm{Data}|\theta)$.
- **How to combine** the two sources of information?

**Bayes' theorem**

# Learning from data - Bayes' theorem

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

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

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

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

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

# Testing for Covid



- A $=\{$Covid$\}$, B $=\{$Positive home test$\}$.
- **Sensitivity**: $96.77\%$. This is $p(B|A) = 0.9677$.
- **Specificity**: $99.20\%$. This is $p(B^c|A^c) = 0.9920$.
- **Prevalence**: $5\%$. This is $p(A) = 0.05$.
- Probability of being sick when test is positive:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A^c)p(A^c)} \approx 0.864.$$

- Probably some symptoms. So maybe $\mathrm{Pr(A)} = 0.7$. Then

$$p(A|B) = 0.9965.$$

- **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** to ensure 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

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

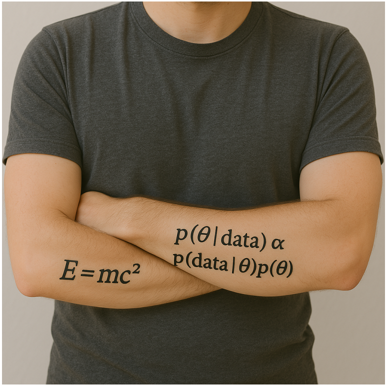
# Great theorems make great tattoos

- 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**

$$x_1, ..., x_n|\theta \overset{iid}{\sim} \mathrm{Bern}(\theta)$$

- **Prior**

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

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

- **Posterior**

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

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

# Beta distribution



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

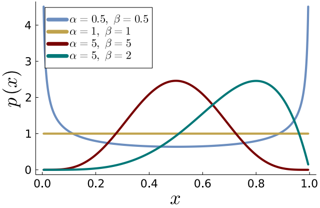
$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

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

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

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

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



# Bayesian updating in Bernoulli trials

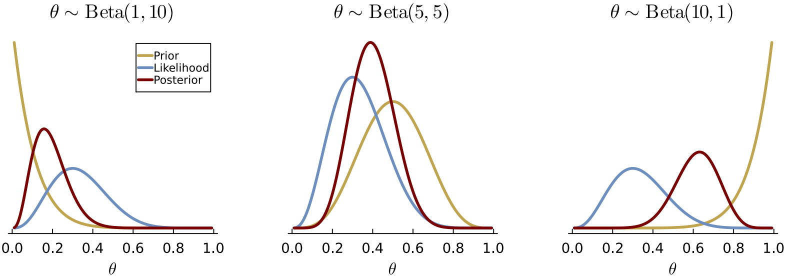
**Conjugate analysis - Bernoulli model**

**Model**: $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Bern}(\theta)$

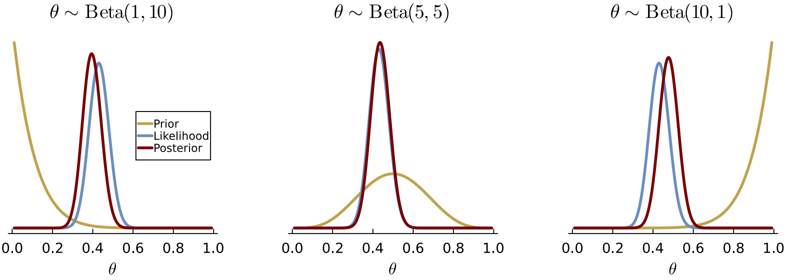
**Prior**: $\theta \sim \mathrm{Beta}(\alpha, \beta)$

**Posterior**: $\theta | x_1, \ldots, x_n \sim \mathrm{Beta}(\alpha + s, \beta + f)$

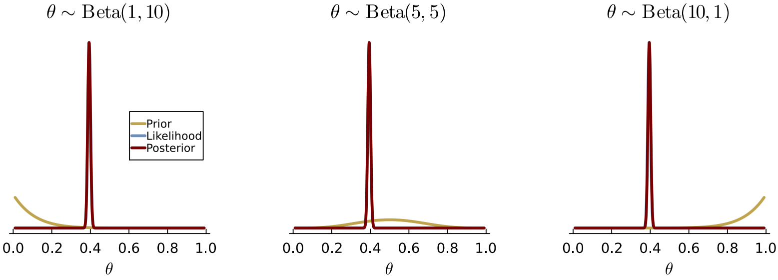
# Spam data (n=10) - Prior is influential



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



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



# Normal data, known variance - uniform prior

- **Model**

$$x_1, ..., x_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2).$$

- **Prior**

$$p(\theta) \propto c \text{ (a constant)}$$

- **Likelihood**

$$\begin{aligned} p(x_1, ..., x_n | \theta, \sigma^2) &= \Pi_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta - \bar{x})^2\right]. \end{aligned}$$

- **Posterior**

$$\theta | x_1, ..., x_n \sim N(\bar{x}, \sigma^2/n)$$

# Normal data, known variance - normal prior

- **Prior**

$$\theta \sim N(\mu_0, \tau_0^2)$$

- **Posterior**

$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\theta, \sigma^2)p(\theta) \\ &\propto N(\theta|\mu_n, \tau_n^2), \end{aligned}$$

where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2},$$

$$\mu_n = w\bar{x} + (1-w)\mu_0,$$

and

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

# Bayesian updating for Normal data

**Conjugate analysis - Gaussian model, known variance**

**Model**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$, $\sigma^2$ known

**Prior**: $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$

**Posterior**: $\theta | x_1, \ldots, x_n \sim \mathrm{N}\big(\mu_n, \tau_n^2\big)$

Posterior precision: $\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$

Posterior mean: $\mu_n = w\bar{x} + (1 - w)\mu_0$

Posterior weight: $w = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau_0^2}$

# Normal data, known variance - normal prior

$$\text{Posterior precision} = \text{Data precision} + \text{Prior precision}$$

$$\text{Posterior mean} =$$

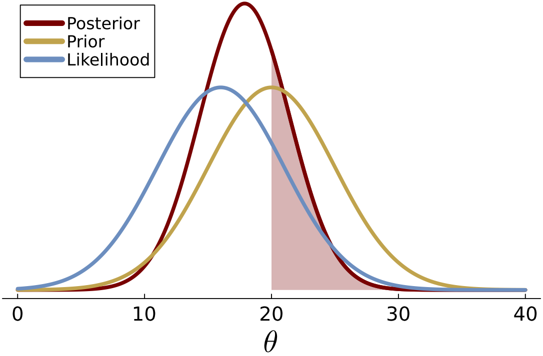
$$\frac{\text{Data precision}}{\text{Posterior precision}}(\text{Data mean}) + \frac{\text{Prior precision}}{\text{Posterior precision}}(\text{Prior mean})$$

# Internet speed

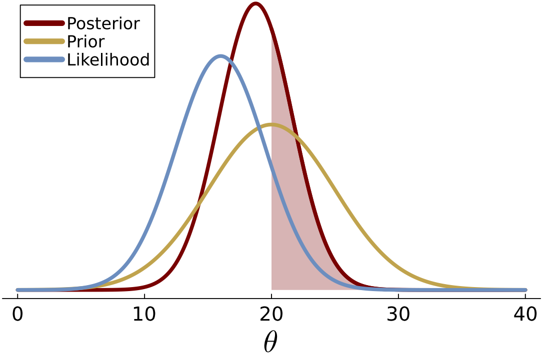


- **Problem**: My internet provider promises an average download speed of at least 20 Mbit/sec. Are they lying?
- **Data**: $\mathrm{x} = (15.77, 20.5, 8.26, 14.37, 21.09)$ Mbit/sec.
- **Model**: $X_1, ..., X_5 \sim N(\theta, \sigma^2)$.
- Assume $\sigma = 5$ (measurements can vary $\pm 10$MBit with 95% probability)
- My **prior**: $\theta \sim N(20, 5^2)$.

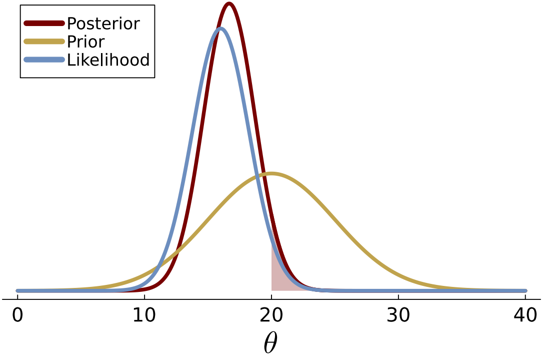
# Internet speed n=1



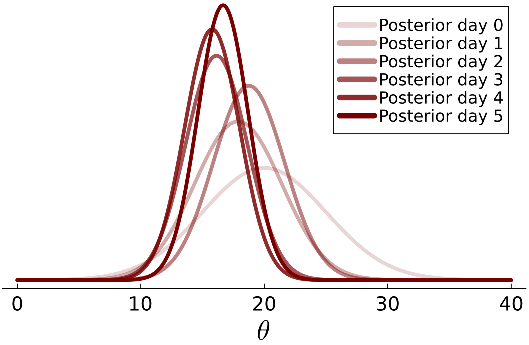
# Internet speed n=2



# Internet speed n=5



# Bayesian updating



# Bayes respects the Likelihood Principle

- **Bernoulli trials with order**:
  $x_1 = 1, x_2 = 0, ..., x_4 = 1, ..., x_n = 1$

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

- **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**.