Bayesian Statistics I

Lecture 1 - The Bayesics, Bernoulli and Normal data

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

mattiasvillani.com @matvil mattiasvillani

Course overview

- Course webpage. Course syllabus.
- Modes of teaching:
  - Lectures (Mattias Villani and Oskar Gustafsson L7-L9)
  - 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
  - Exam: Pen and paper + Computer

Lecture overview

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

Likelihood function - Bernoulli trials

- **Bernoulli trials**:

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

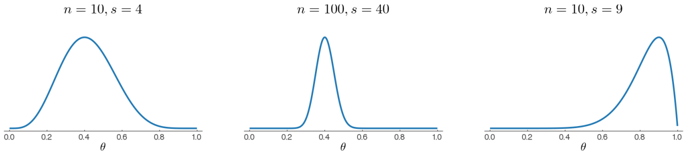
- **Likelihood** from $s = \sum_{i=1}^{n} x_i$ successes and $f = n - s$ failures.

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

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

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

$n = 10, s = 4$ $\qquad$ $n = 100, s = 40$ $\qquad$ $n = 10, s = 9$



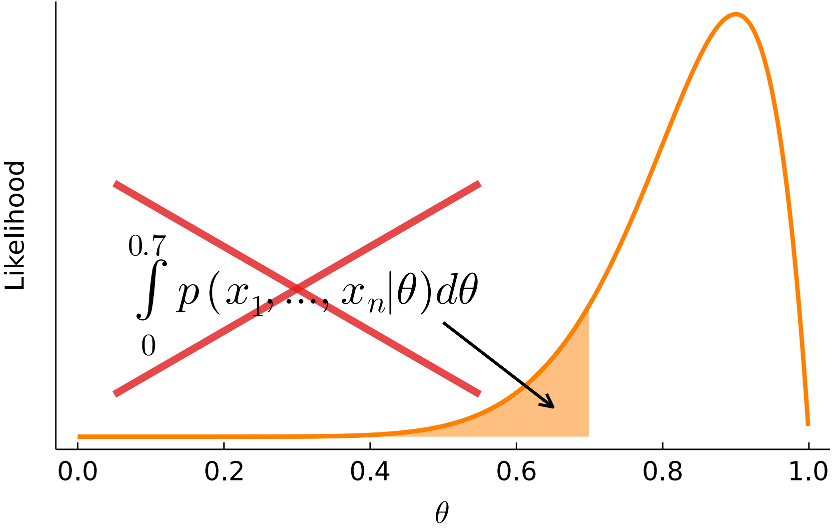
The likelihood function

- Say it out loud:

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

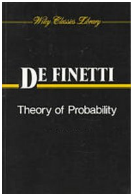
- The symbol $p(x_1, ..., x_n|\theta)$ plays two different roles:
- **Probability distribution** for the data.
  - The data $\mathrm{x} = (x_1, ..., x_n)$ are random.
  - θ is fixed.
- **Likelihood function** for the parameter
  - The data $\mathrm{x} = (x_1, ..., x_n)$ are fixed.
  - $p(x_1, ..., x_n|\theta)$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

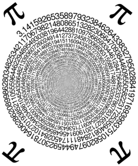
Probabilities from the likelihood?



Uncertainty and subjective probability

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





Bayesian learning

- **Bayesian learning** about a model parameter θ:
  - state your **prior** knowledge as a probability distribution $p(\theta)$.
  - collect **data** Data and form the **likelihood** function $p(\mathrm{Data}|\theta)$.
  - **combine** prior knowledge $p(\theta)$ with 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 θ

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

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

Medical diagnosis

- A = {Very rare disease}, B ={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

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

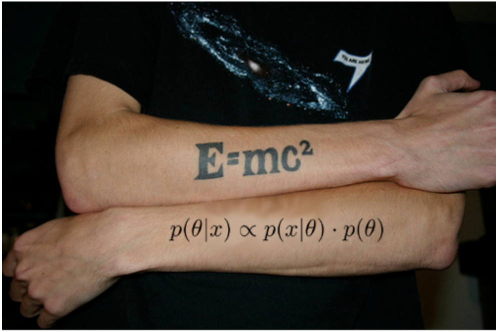
Great theorems make great tattoos

- 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**

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

- **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**

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

- Posterior is proportional to the $\mathrm{Beta}(\alpha+s, \beta+f)$ density.
- The prior-to-posterior mapping:

$$\theta \sim \mathrm{Beta}(\alpha, \beta) \overset{x_1, ..., x_n}{\Longrightarrow} \theta | x_1, ..., x_n \sim \mathrm{Beta}(\alpha+s, \beta+f)$$

Beta distribution

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

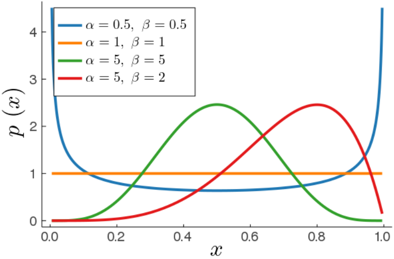
$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

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

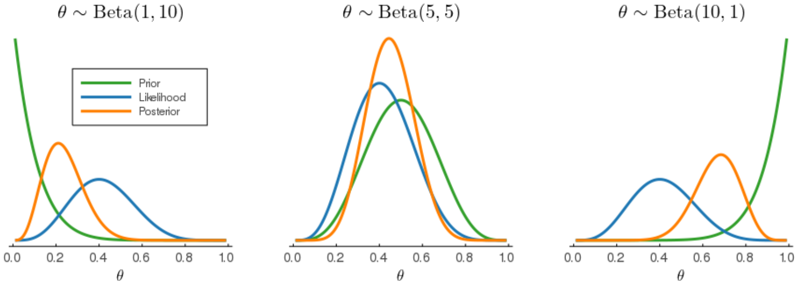
$$\mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

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

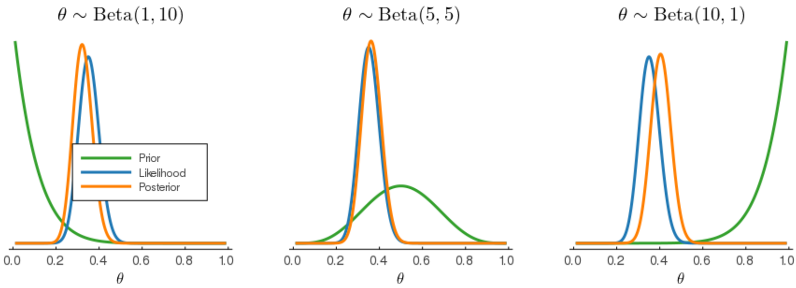
$\Gamma(\alpha)$ is the Gamma function.



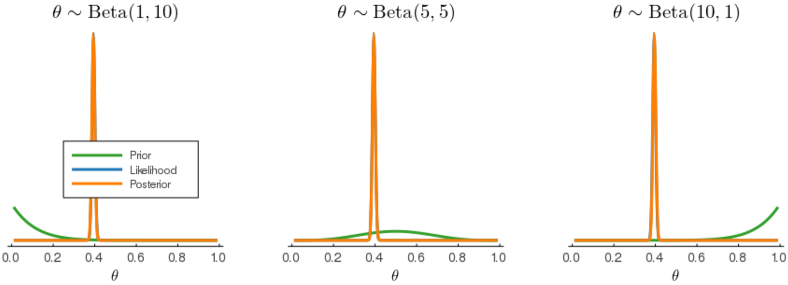
Spam data (n=10) - Prior is influential



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**:
  $x_1 = 1, x_2 = 0, ..., x_4 = 1, ..., x_n = 1$

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

- **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**.

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

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

Normal data, known variance - normal prior

$$\theta \sim N(\mu_0, \tau_0^2) \overset{x_1,\ldots,x_n}{\Longrightarrow} \theta|x \sim N(\mu_n, \tau_n^2).$$

Posterior precision = Data precision + Prior precision

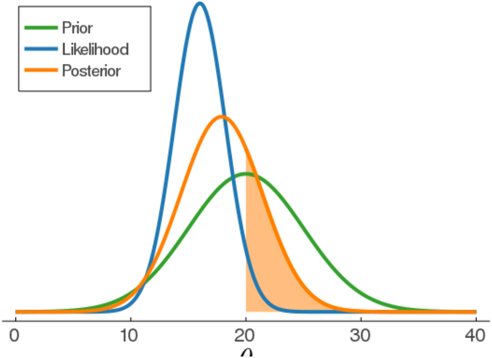
Posterior mean =

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

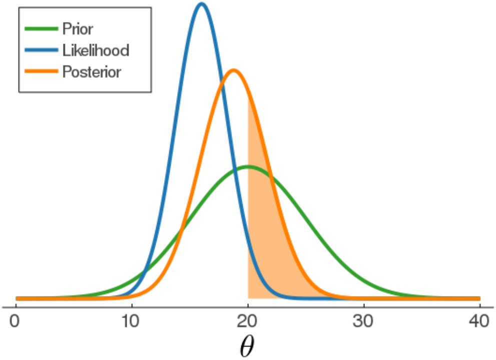
Download speed

- **Problem**: My internet provider promises an average download speed of at least 20 Mbit/sec. Are they lying?
- **Data**: $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- **Model**: $X_1, ..., X_5 \sim N(\theta, \sigma^2)$.
- Assume $\sigma = 5$ (measurements can vary $\pm$10MBit with 95% probability)
- My **prior**: $\theta \sim N(20, 5^2)$.

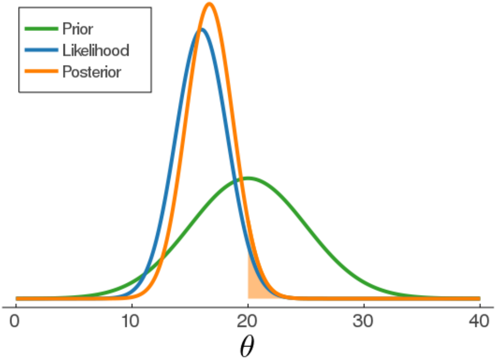
Download speed n=1



Download speed n=2



Download speed n=5



Bayesian updating