

BAYESIAN STATISTICS - LECTURE 1

LECTURE 1: LIKELIHOOD. BAYESICS. BERNOULLI

MATTIAS VILLANI

DEPARTMENT OF STATISTICS

STOCKHOLM UNIVERSITY

AND

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE

LINKÖPING UNIVERSITY

- Course **webpage** is **here**.
- Course **syllabus** is **here**.
- Modes of teaching:
 - **Lectures** (**Mattias Villani**)
 - **Mathematical exercises** (**Munezero Parfait** and **Oscar Oelrich**)
 - **Computer labs** (**Munezero Parfait** and **Oscar Oelrich**)
- **Modules:**
 - The **Bayesics**, single- and multiparameter models
 - **Regression** and **Classification models**
 - **Advanced models** and **Posterior Approximation** methods
 - **Model Inference, Model evaluation** and **Variable Selection**
- **Examination**
 - Lab reports
 - Home exam

- The **likelihood function**
- **Bayesian inference**
- **Bernoulli model**

THE LIKELIHOOD FUNCTION - BERNOULLI TRIALS

■ Bernoulli trials:

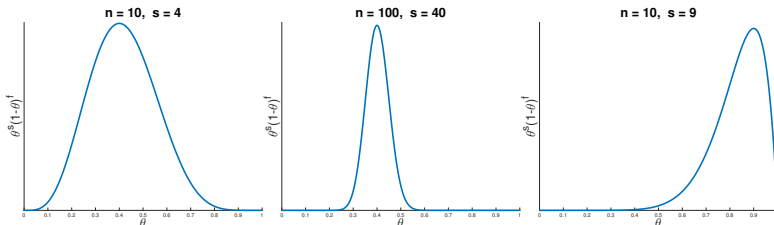
$$X_1, \dots, X_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta).$$

- **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 θ .



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.

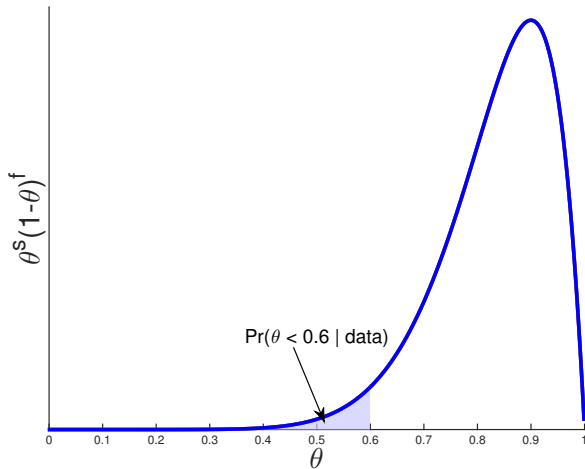
- The data $\mathbf{x} = (x_1, \dots, x_n)$ are random.
- θ is fixed.

■ **Likelihood function** for the parameter

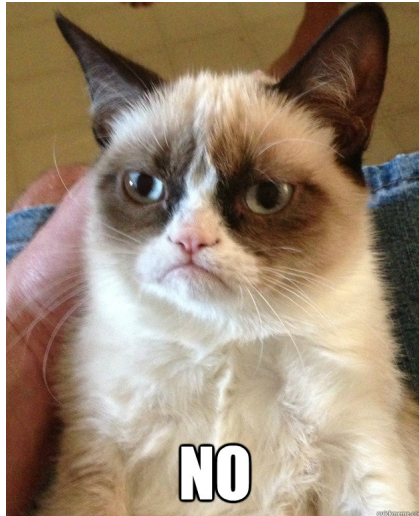
- The data $\mathbf{x} = (x_1, \dots, x_n)$ are fixed.
- $p(x_1, \dots, x_n | \theta)$ is function of θ .

PROBABILITIES FROM THE LIKELIHOOD!!

n = 10, s = 9

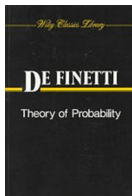


PROBABILITIES FROM THE LIKELIHOOD!!



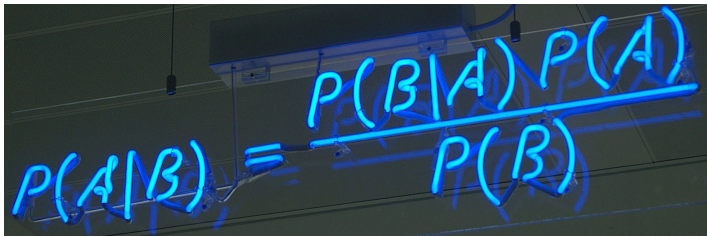
UNCERTAINTY AND SUBJECTIVE PROBABILITY

- $\Pr(\theta < 0.6 | \text{data})$ 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 $\Pr(10\text{th decimal of } \pi = 9) = 0.1$ makes sense.



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

Bayes' theorem


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

- 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 θ

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

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

BAYES' THEOREM FOR 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

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

- The integral $p(Data) = \int_{\theta} p(Data|\theta)p(\theta)d\theta$ can make you cry.
- **$p(Data)$ is just a constant** so that $p(\theta|Data)$ integrates to one.
- 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

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

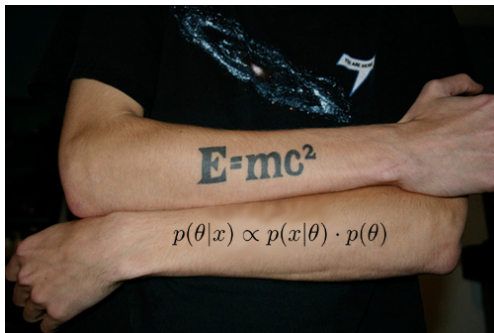
GREAT THEOREMS MAKE GREAT TATTOOS

- All you need to know:

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

or

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



■ Model

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

■ Prior

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

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

■ Posterior

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

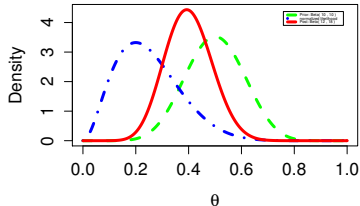
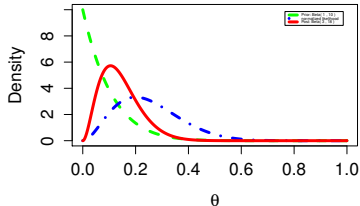
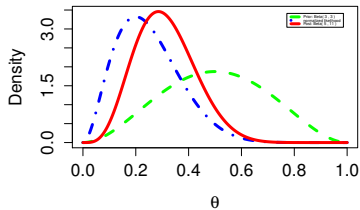
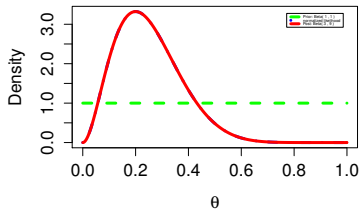
- Posterior is proportional to the $\text{Beta}(\alpha + s, \beta + f)$ density.
- The prior-to-posterior mapping:

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

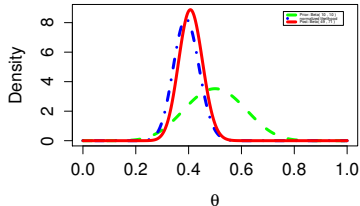
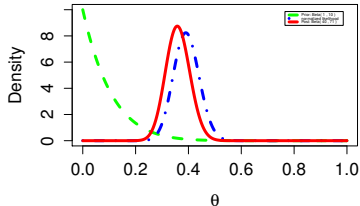
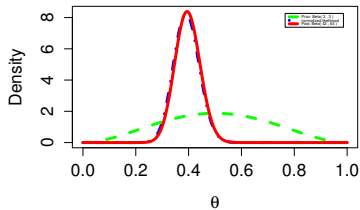
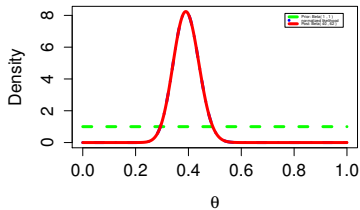
BERNOULLI EXAMPLE: SPAM EMAILS

- George has gone through his collection of 4601 e-mails.
- He classified 1813 of them to be spam.
- Let $x_i = 1$ if i :th email is spam.
- **Model:** $x_i | \theta \stackrel{iid}{\sim} \text{Bern}(\theta)$
- **Prior:** $\theta \sim \text{Beta}(\alpha, \beta)$.
- **Posterior**
 $\theta | \mathbf{x} \sim \text{Beta}(\alpha + 1813, \beta + 2788)$

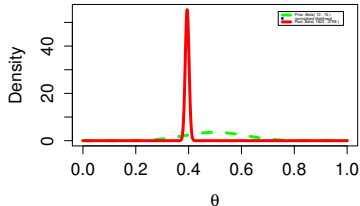
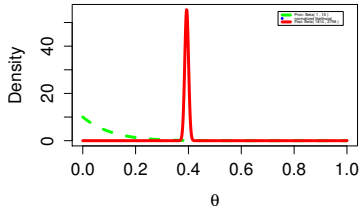
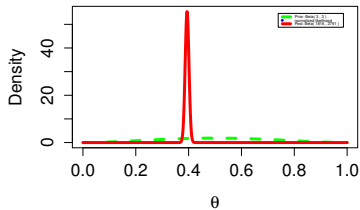
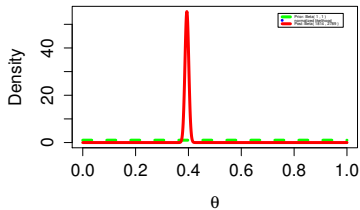
SPAM DATA (N=10): PRIOR SENSITIVITY



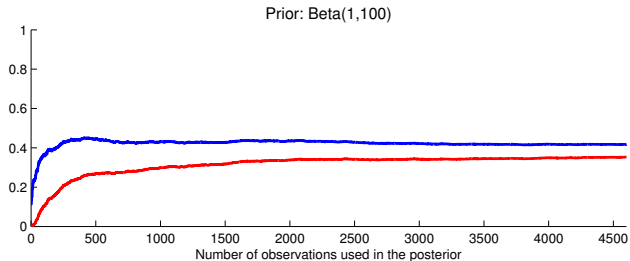
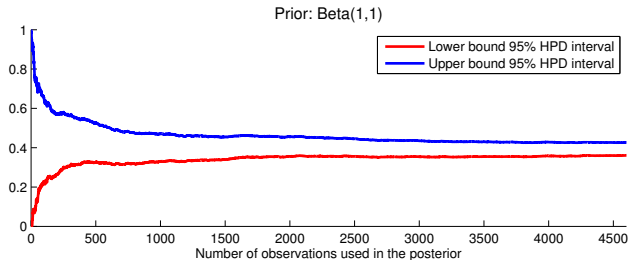
SPAM DATA (N=100): PRIOR SENSITIVITY



SPAM DATA (N=4601): PRIOR SENSITIVITY



SPAM DATA: POSTERIOR CONVERGENCE



THREE SHADES OF BINARY - A SINGLE SHADE OF BAYES

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

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

- **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**.