Workshop: Intro to Bayesian Learning
Lecture 1 - The Bayesics

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- The likelihood function
- Bayes and subjective probability
- Bayesian analysis of Bernoulli data
- Bayesian analysis of Gaussian data with known variance
- Bayesian analysis of Poisson counts

Likelihood function - Bernoulli trials

- **Bernoulli trials**:

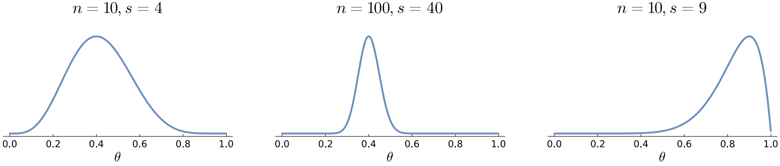
$$X_1, ..., X_n|\theta \overset{iid}{\sim} Bern(\theta).$$

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

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

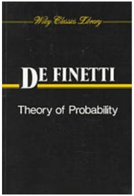
- **Maximum likelihood estimator** $\hat{\theta}$ maximizes $p(x_1, ..., x_n|\theta)$.

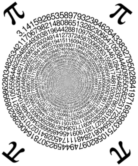
- Given the data $x_1, ..., x_n$, plot $p(x_1, ..., x_n|\theta)$ as a function of θ.



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.





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

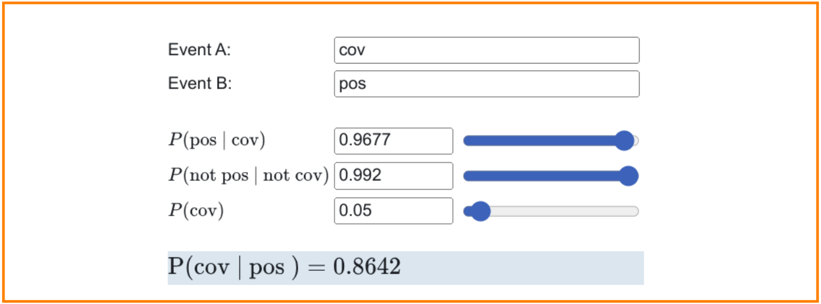
$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

- 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**.

Bayes' theorem for Covid tests



Great theorems make great tattoos

- Bayes theorem

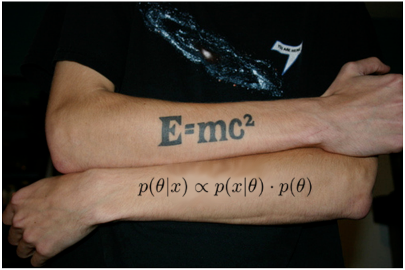
$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)} = \frac{p(Data|\theta)p(\theta)}{\int_\theta p(Data|\theta)p(\theta)d\theta}.$$

- 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**

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

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

- **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

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

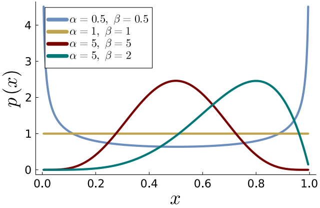
$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

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

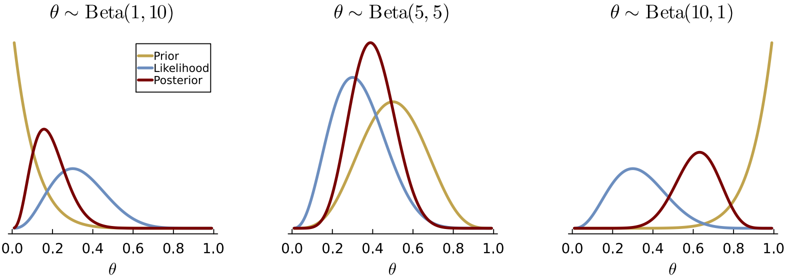
$$\mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

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

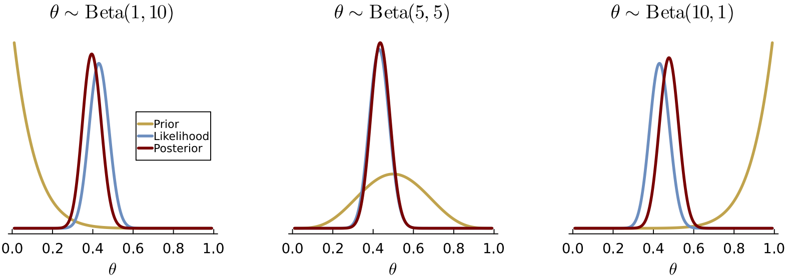
$\Gamma(\alpha)$ is the Gamma function.



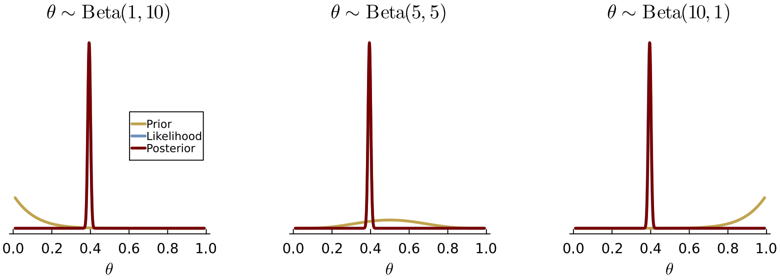
Spam data (n=10) - Prior is influential



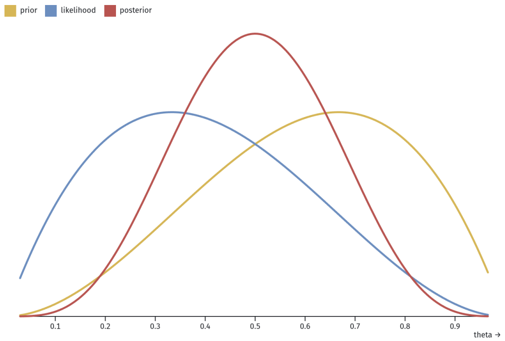
Spam data (n=100) - Prior is less influential



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



👀 Bernoulli model - Beta prior



Prior-to-Posterior mapping. The likelihood is normalized.

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

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

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

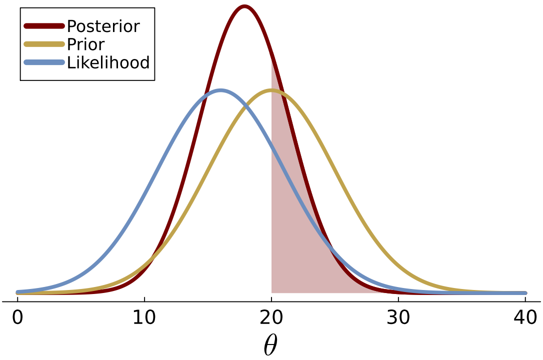
$$\text{Posterior mean} =$$

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

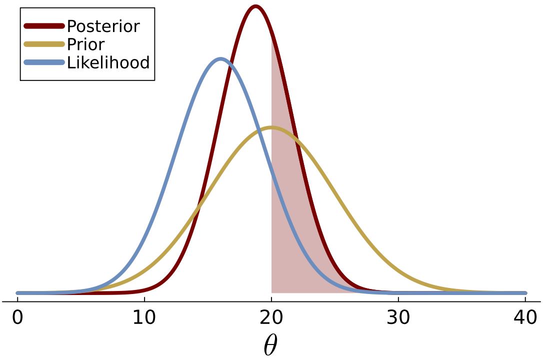
Internet speed

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

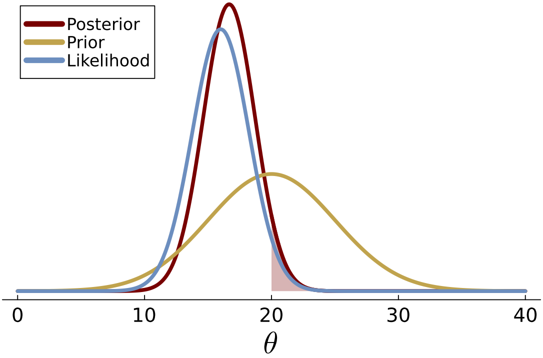
Internet speed n=1



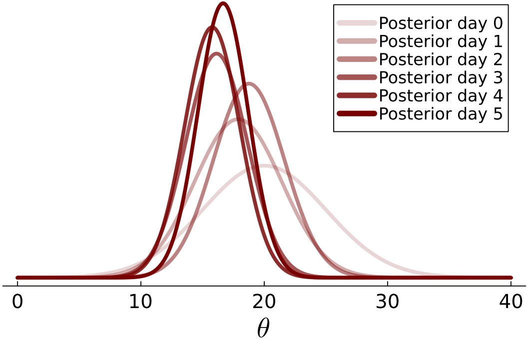
Internet speed n=2



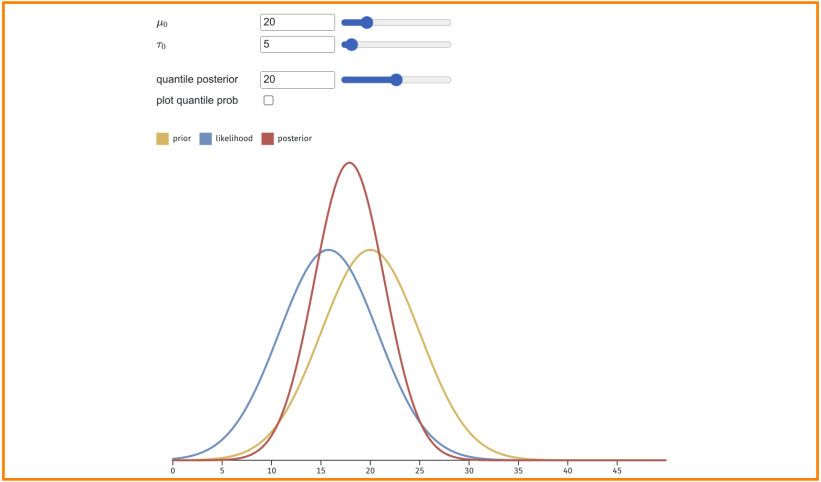
Internet speed n=5



Bayesian updating



Normal model known variance - normal prior



Poisson model

- **Model**

$$y_1, ..., y_n|\theta \overset{iid}{\sim} \mathrm{Pois}(\theta)$$

- **Poisson distribution**

$$p(y) = \frac{\theta^y e^{-\theta}}{y!}$$

- **Likelihood** from iid Poisson sample $y = (y_1, ..., y_n)$

$$p(y|\theta) = \left[\prod_{i=1}^{n} p(y_i|\theta)\right] \propto \theta^{(\sum_{i=1}^{n} y_i)} \exp(-\theta n),$$

- **Prior**

$$p(\theta) \propto \theta^{\alpha-1} \exp(-\theta\beta) \propto \mathrm{Gamma}(\alpha, \beta)$$

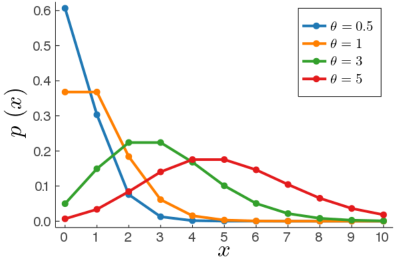
Poisson distribution

$X \sim \mathrm{Pois}(\theta)$
for $X \in 0, 1, 2, \ldots$

$$p(x) = \frac{\theta^x e^{-\theta}}{x!}$$

$$\mathbb{E}(X) = \theta$$

$$\mathbb{V}(X) = \theta$$



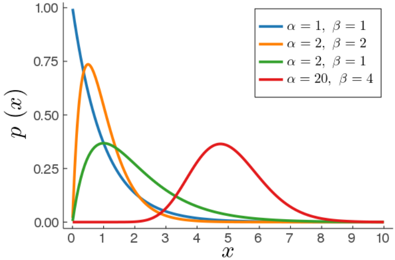
Gamma distribution

$$X \sim \mathrm{Gamma}(\alpha, \beta)$$

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

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

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



Poisson posterior

- **Posterior**

$$p(\theta|y_1, ..., y_n) \propto \left[\prod_{i=1}^{n} p(y_i|\theta)\right] p(\theta)$$
$$\propto \theta^{\sum_{i=1}^{n} y_i} \exp(-\theta n)\theta^{\alpha-1} \exp(-\theta\beta)$$
$$= \theta^{\alpha+\sum_{i=1}^{n} y_i - 1} \exp[-\theta(\beta + n)],$$

which is proportional to $\mathrm{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$.

- **Prior-to-Posterior mapping**

Model: $y_1, ..., y_n|\theta \overset{iid}{\sim} \mathrm{Pois}(\theta)$

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

Posterior: $\theta|y_1, ..., y_n \sim \mathrm{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$.

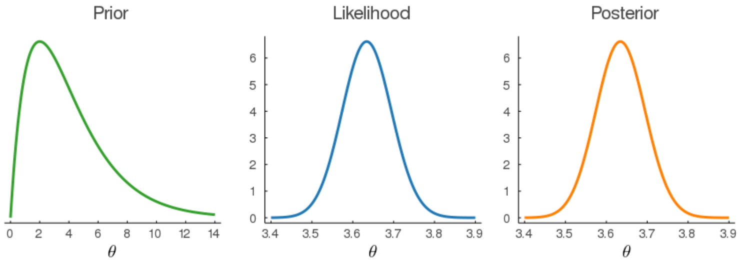
Example - Number of bids in eBay auctions

- **Data**:
  - Number of placed bids in $n = 1000$ eBay coin auctions.
  - Sum of counts: $\sum_{i=1}^{n} y_i = 3635$.
  - Average number bids per auction: $\bar{y} = 3635/1000 = 3.635$.
- **Prior**: $\theta \sim Gamma(\alpha, \beta)$ with $\alpha = 2$, $\beta = 1/2$.

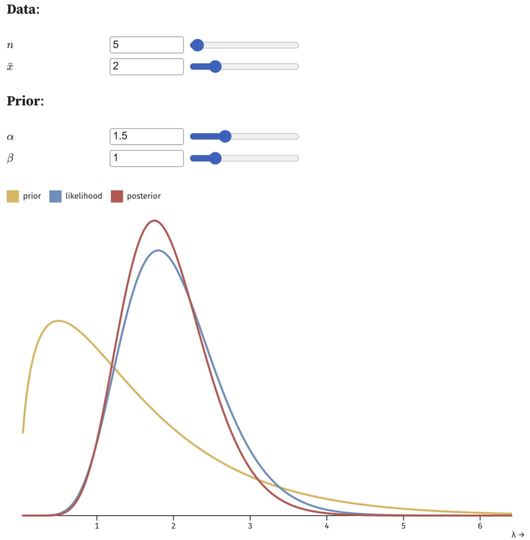
$$E(\theta) = \frac{\alpha}{\beta} = 4$$

$$SD(\theta) = \frac{\alpha}{\beta^2} = 2.823$$

eBay data - Posterior of θ



Poisson model - Gamma prior



Bayesian point estimates

- Bayes gives a whole **posterior distribution**. Contains all info.
- Sometimes convenient to summarize by a **point estimate**
  - Posterior **mean** (quadratic loss)
  - Posterior **median** (linear loss)
  - Posterior **mode** (zero-one loss)
- A 95% **posterior credible interval** $[l, u]$ for a parameter θ conditional on data $\boldsymbol{x}$ satisfies

$$\Pr(\mathrm{l} \leq \theta \leq \mathrm{u} | \boldsymbol{x}) = 0.95$$

- Frequentist interval: a random interval that covers the true θ in 95% of all datasets drawn from the population. 😢

Bayesian credible intervals

distribution: binormal mixture

location μ_1 -2

location μ_2 2

scale σ_1 1

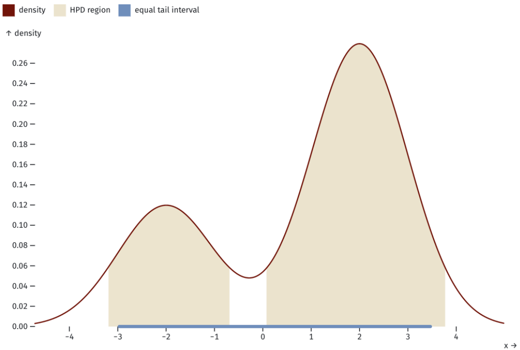
scale σ_2 1

weight w_1 0.3

coverage % 90

The 90% HPD region is (-3.191, -0.689) ∪ (0.077, 3.771)

The 90% equal tail region is (-2.967, 3.465)



Frequentist coverage of credible intervals

confidence level 90

sample size, n 50

lower limit in plot 0

plot interval type: ☑ wald ☑ wilson ☑ bayes

Average absolute deviation to target coverage:

Wald: 0.0464 | Wilson: 0.0157 | Bayes: 0.0171