Bayesian Inference

PhD course in Statistical Inference

Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

Overview of the day

- **Introduction to subjective probability and Bayesian inference**
- **Prediction, Decisions, Exponential family**
- **Bayesian large sample theory and posterior approximation**
- **Bayesian computations**
- **Bayesian model comparison**

- **Subjective probability**
- **The Bayesics**
- **Bayes for the Normal model**
- **Priors**

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

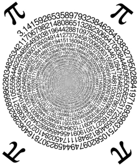
- Likelihood function is **NOT** a probability distribution for θ.

- Statements like $\Pr(\theta > c)$ makes no sense.

- Unless ...

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.

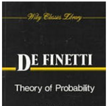


Uncertainty and subjective probability

"The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense determined, or known by other people, and so on, is of no consequence." - Bruno de Finetti

*"**Probability does not exist**" - Bruno de Finetti in the introduction to his classic book A Theory of Probability*

- Subjective probability applies also to **non-repeatable experiments**.
- Subjective probabilities must satisfy the usual **axioms for probabilities**. Dutch book arguments. Axiomatic theory.



- **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

- 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**.

- Bayes theorem

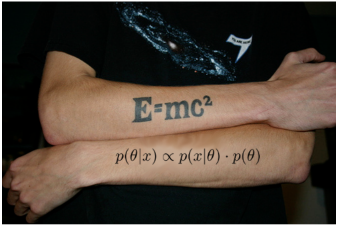
$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)}$$

- All you need to know:

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

or

Posterior $\propto$ Likelihood $\cdot$ Prior



Normal data, known variance - uniform prior

- **Model**

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

- **Prior**

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

- **Likelihood**

$$p(x_1, \ldots, x_n | \theta, \sigma^2) \quad \propto \quad \exp\left[-\frac{1}{2(\sigma^2/n)} (\theta - \bar{x})^2 \right]$$

- **Posterior**

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

- **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

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

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

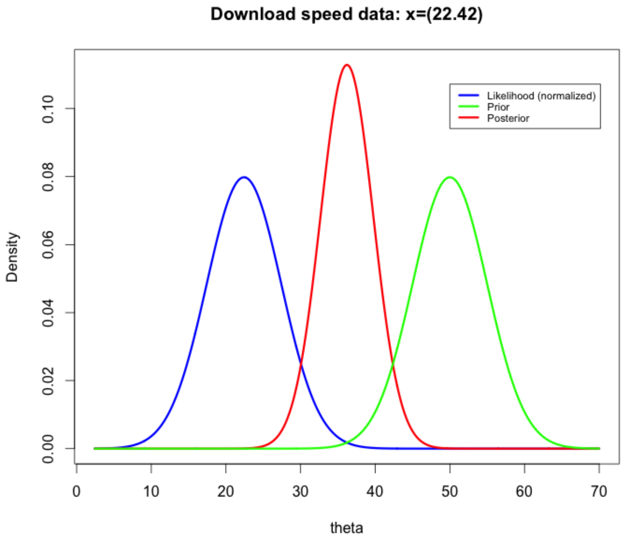
and

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

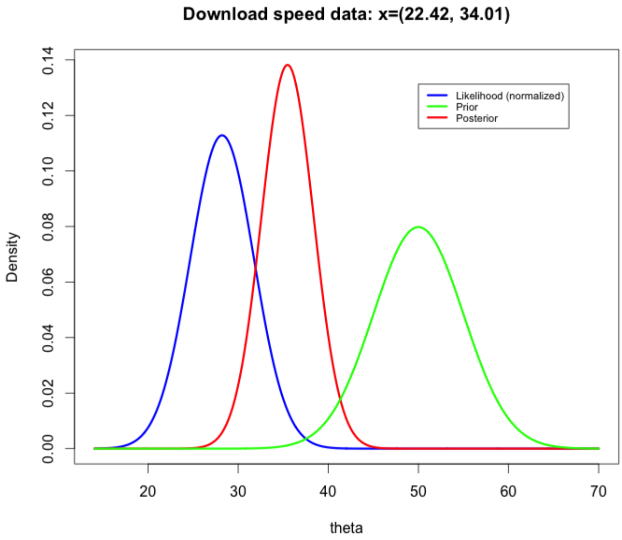
- Proof: complete the squares in the exponential.

- **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**: $\theta \sim N(50, 5^2)$.

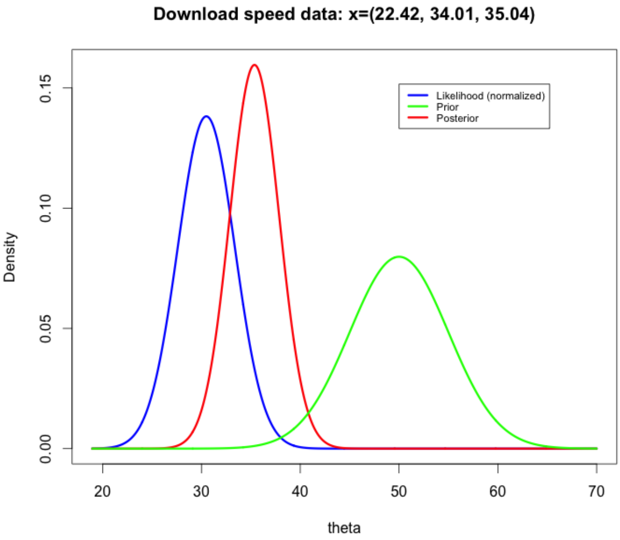
Download speed n=1



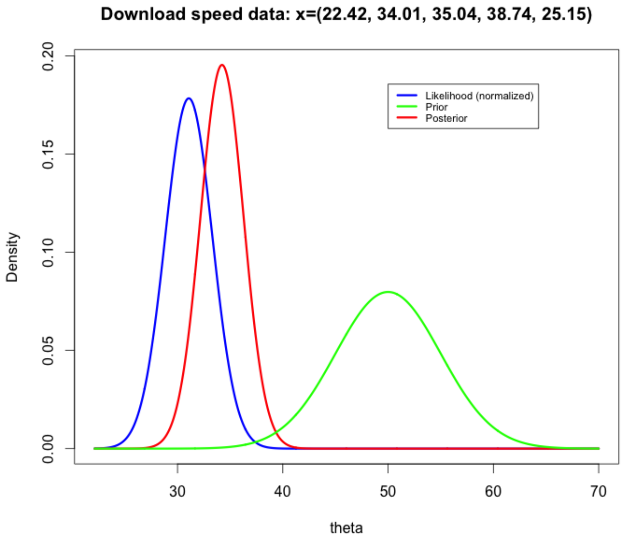
Download speed n=2



DOWNLOAD SPEED N=3



Download speed n=5



- Models with **multiple parameters** $\theta_1, \theta_2, \ldots$.
- Examples: $x_i \overset{iid}{\sim} N(\theta, \sigma^2)$; multiple regression ...
- **Joint posterior distribution**

$$p(\theta_1, \theta_2, ..., \theta_p|y) \propto p(y|\theta_1, \theta_2, ..., \theta_p)p(\theta_1, \theta_2, ..., \theta_p).$$

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

- **Marginalize** out parameter of no direct interest (**nuisance**).
- Example: $\theta = (\theta_1, \theta_2)'$. **Marginal posterior** of θ_1

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2 = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2.$$

- **Model**

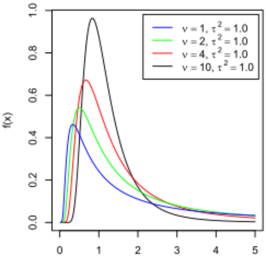
$$y_1, \ldots, y_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

- **Conjugate prior**

$$\theta | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

$$\sigma^2 \sim Inv\text{-}\chi^2(\nu_0, \sigma_0^2)$$

- **Scaled inverse χ^2 distribution**



- **Posterior**

$$\theta|\mathbf{y},\sigma^2 \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$
$$\sigma^2|\mathbf{y} \sim \textit{Inv-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

- **Marginal posterior**

$$\theta|\mathbf{y} \sim t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right)$$

- **Binomial sampling**:

$$s|\theta \stackrel{iid}{\sim} Bin(n, \theta), \; n \text{ fixed}.$$

- Prior

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

- Posterior

$$\begin{aligned} p(\theta|s) &\propto p(s|\theta)p(\theta) \\ &= \binom{n}{s}\theta^s(1-\theta)^f \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1} \end{aligned}$$

so, $\theta|s \sim \mathrm{Beta}(\alpha+s, \beta+f)$.

- **Negative binomial sampling**:

$$n|\theta \overset{iid}{\sim} Negbin(s, \theta), \ s \text{ fixed}.$$

- Posterior

$$\begin{aligned} p(\theta|n) &\propto p(n|\theta)p(\theta) \\ &= \binom{n-1}{s-1}\theta^s(1-\theta)^f \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1} \end{aligned}$$

  so again, $\theta|s \sim \mathrm{Beta}(\alpha+s, \beta+f)$.
- Same posterior regardless of how the data was obtained.
- Bayesian inference respects the **likelihood principle**:

$$p(\theta|\mathbf{x}) = \frac{c \cdot p(\mathbf{x}|\theta)p(\theta)}{\int c \cdot p(\mathbf{x}|\theta)p(\theta)d\theta} = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}$$

  for any $c > 0$.

- The prior should be determined (**elicited**) by an **expert**. Typically, expert$\neq$statistician.
- Elicit the prior on **a quantity that the expert knows well**. Convert afterwards.
- **Ask probabilistic questions** to the expert:
  - $E(\theta) = ?$
  - $SD(\theta) = ?$
  - $Pr(\theta < c) = ?$
  - $Pr(y > c) = ?$
- **Show some consequences** of the elicitated prior to the expert.
- Beware of **psychological effects**, such as anchoring.

- **Autoregressive process** or order p

$$y_t \;=\; \mu + \phi_1(y_{t-1} - \mu) + \ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \;\; \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

- Informative prior on the unconditional mean: $\mu \sim N(\mu_0, \tau_0^2)$.
- “Noninformative” prior on σ^2: $p(\sigma^2) \propto 1/\sigma^2$
- Assume $\phi_i \sim N(\mu_i, \psi_i), i = 1, \ldots, p$ are independent a priori.
- Prior on $\phi = (\phi_1, \ldots, \phi_p)$ centered on persistent AR(1) process: $\mu_1 = 0.8, \mu_2 = \ldots = \mu_p = 0$
- $Var(\phi_i) = \frac{c}{i^\lambda}$. “Longer” lags are more likely to be zero a priori.

Different types of prior information

- Real **expert information**. Combo of previous studies and experience.
- **Vague prior** information.
- **Reporting priors**. Easy to understand the information they contain.
- **Smoothness priors**. Regularization. Shrinkage. Big thing in modern statistics/machine learning.

- **Observed information**

$$J_{\theta,\mathbf{x}} = -\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial\theta^2}\Big|_{\theta=\hat{\theta}}$$

- **Fisher information**

$$I_\theta = E_{\mathbf{x}|\theta}\left(J_{\theta,\mathbf{x}}\right)$$

- A common non-informative prior is **Jeffreys' prior**

$$p(\theta) = \left|I_\theta\right|^{1/2}.$$

- **Invariant** to 1:1 transformations of θ.
- Often non-conjugate.
- Often problematic in multiparameter settings.

$$x_1, \ldots, x_n|\theta \stackrel{iid}{\sim} Bern(\theta).$$

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

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

$$\frac{d^2 \ln p(\mathbf{x}|\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{f}{(1-\theta)^2}$$

$$I(\theta) = \frac{E_{\mathbf{x}|\theta}(s)}{\theta^2} + \frac{E_{\mathbf{x}|\theta}(f)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Thus, the Jeffreys' prior is

$$p(\theta) = |I(\theta)|^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto Beta(1/2, 1/2).$$

- Jeffreys' prior:

$$n|\theta \overset{iid}{\sim} NegBin(s, \theta).$$

$$\ln p(\mathbf{x}|\theta) = \ln \binom{n-1}{s-1} + s \ln \theta + f \ln(1-\theta)$$

$$\frac{d^2 \ln p(\mathbf{x}|\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{f}{(1-\theta)^2}$$

$$I(\theta) = \frac{s}{\theta^2} + \frac{E_{n|\theta}(n-s)}{(1-\theta)^2} = \frac{s}{\theta^2} + \frac{s/\theta - s}{(1-\theta)^2} = \frac{s}{\theta^2(1-\theta)}$$

- Thus, the Jeffreys' prior is

$$p(\theta) = |I(\theta)|^{1/2} \propto \theta^{-1}(1-\theta)^{-1/2} \propto Beta(\theta|0, 1/2).$$

- Jeffreys' prior is **improper**, but the posterior is proper: $\theta|n \sim \mathrm{Beta}(s, f + 1/2)$ which is proper since $s \geq 1$.
- Jeffreys' prior **violates the likelihood principle** because $I(\theta)$ is sampling-based.

- **Maximum entropy prior**: choose prior with maximum entropy (most uncertain). Problematic for continuous parameters.
- **Reference prior**: Choose the prior that maximizes the expected value of perfect information about θ.
- Under the conditions that guarantees asymptotic normality of posterior: Reference = Jeffreys.