

Bayesian Learning

Lecture 2 - Normal and Poisson data. Prior elicitation.

Bertil Wegmann

Department of Computer and Information Science
Linköping University



Lecture overview

- The **Normal model** with known variance
- The **Poisson model**
- **Conjugate priors**
- **Prior elicitation**
- **Jeffreys' prior**

Normal data, known variance - uniform prior

■ Model

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

■ Prior

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

■ Likelihood

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

■ Posterior

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

Normal data, known variance - normal prior

■ Prior

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

■ Posterior

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

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

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

Posterior precision = Data precision + Prior precision

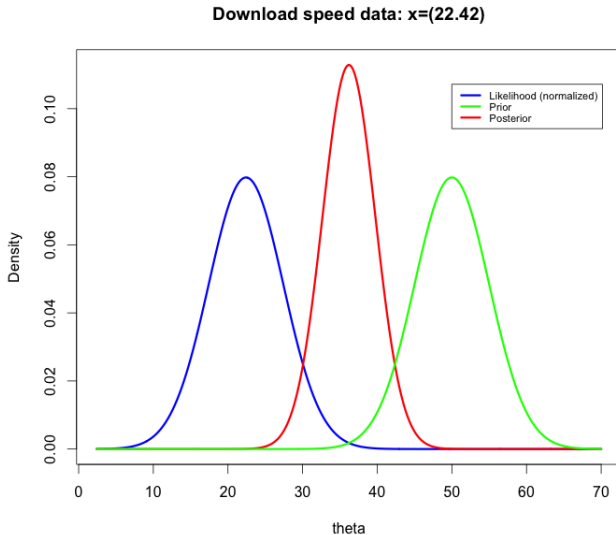
Posterior mean =

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

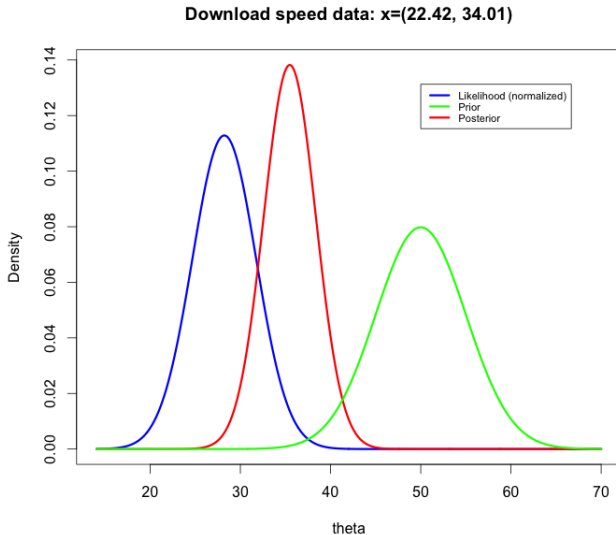
Download speed

- **Data:** $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- **Model:** $X_1, \dots, X_5 | \theta, \sigma^2 \stackrel{iid}{\sim} N(\theta, \sigma^2)$.
- Assume $\sigma = 5$ (measurements can vary ± 10 MBit with 95% probability)
- My **prior:** $\theta \sim N(50, 5^2)$.

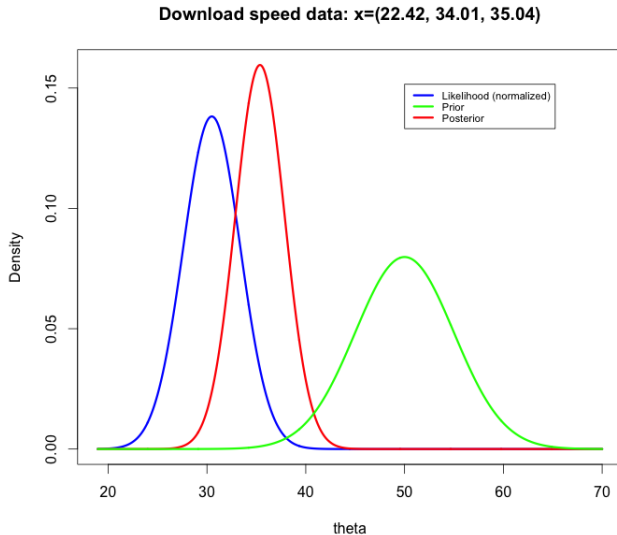
Download speed $n=1$



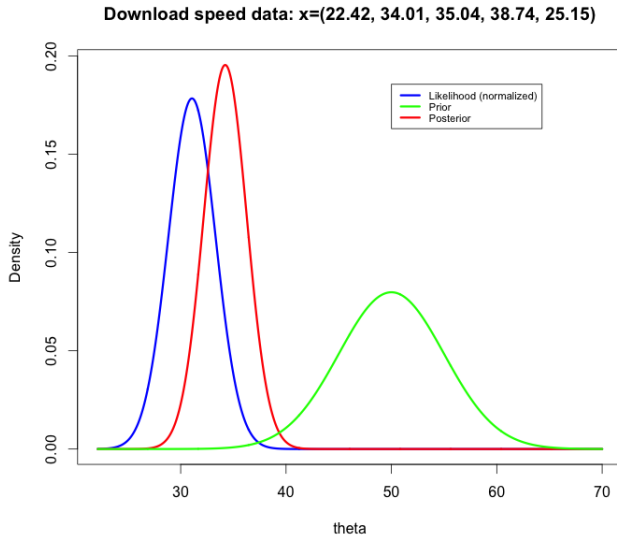
Download speed $n=2$



Download speed $n=3$

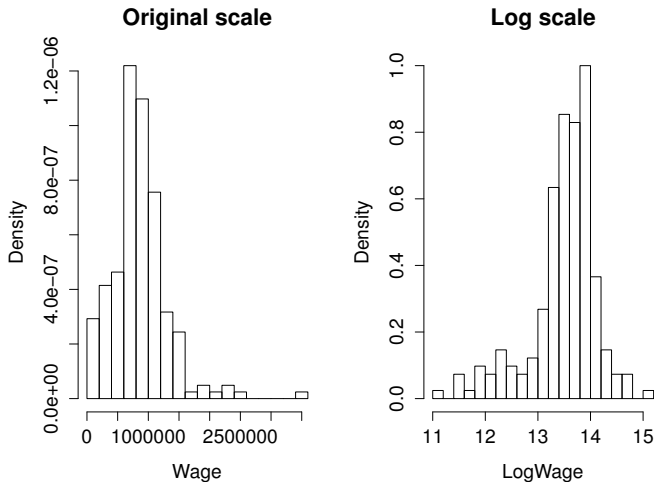


Download speed $n=5$



Canadian wages data

- Data on wages for 205 Canadian workers.



Canadian wages

■ Model

$$X_1, \dots, X_n | \theta \sim N(\theta, \sigma^2), \sigma^2 = 0.4$$

■ Prior

$$\theta \sim N(\mu_0, \tau_0^2), \mu_0 = 12 \text{ and } \tau_0 = 10$$

■ Posterior

$$\theta | x_1, \dots, x_n \sim N(\mu_n, \tau_n^2),$$

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

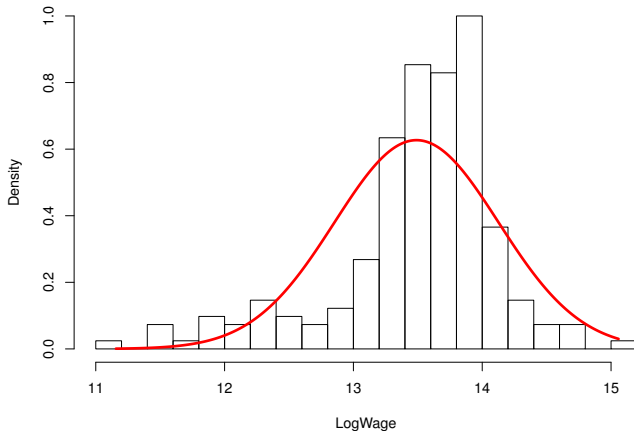
■ For the Canadian wage data:

$$w = \frac{\sigma^{-2}n}{\sigma^{-2}n + \tau_0^{-2}} = \frac{2.5 \cdot 205}{2.5 \cdot 205 + 1/100} = 0.999.$$

$$\mu_n = w\bar{x} + (1 - w)\mu_0 = 0.999 \cdot 13.489 + (1 - 0.999) \cdot 12 \approx 13.489$$

$$\tau_n^2 = (2.5 \cdot 205 + 1/100)^{-1} = 0.00195$$

Canadian wages data - model fit



Poisson model

■ Model

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Pois}(\theta)$$

■ Poisson distribution

$$p(y_i | \theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}, \quad i = 1, \dots, n$$

■ Likelihood from iid Poisson sample $y = (y_1, \dots, y_n)$

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

■ Prior

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

which contains the info: $\alpha - 1$ counts in β observations.

Poisson model, cont.

■ Posterior

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

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

■ Prior-to-Posterior mapping

Model: $y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Pois}(\theta)$

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

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

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 of bids per auction: $\bar{y} = 3635/1000 = 3.635$.

■ Prior: $\alpha = 2, \beta = 1/2$.

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

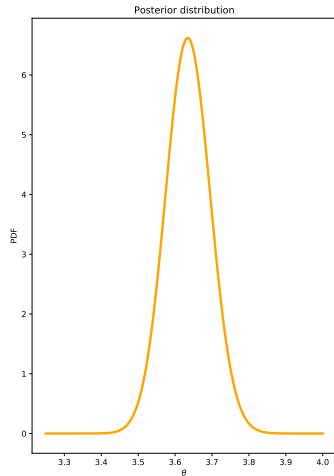
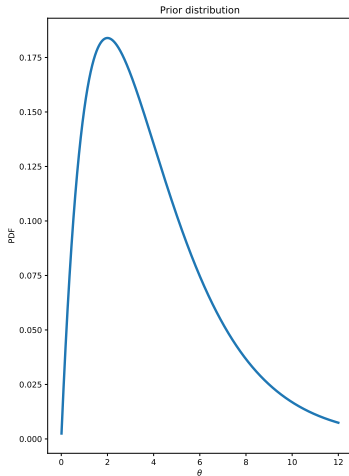
$$SD(\theta) = \left(\frac{\alpha}{\beta^2} \right)^{1/2} = 2.823$$

■ Posterior

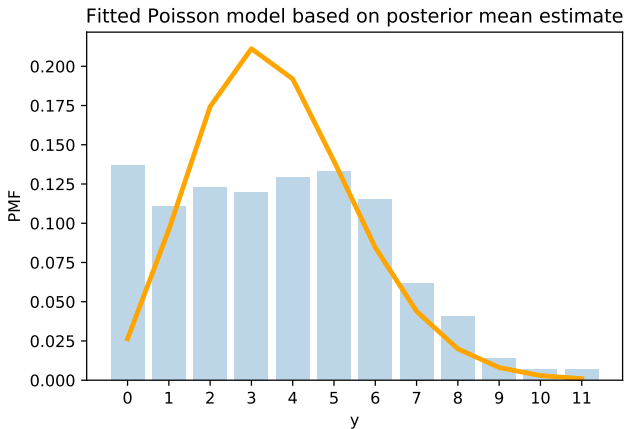
$$E(\theta|\mathbf{y}) = \frac{\alpha + \sum_{i=1}^n y_i}{\beta + n} = \frac{2 + 3635}{1/2 + 1000} \approx 3.635.$$

$$SD(\theta|\mathbf{y}) = \left(\frac{\alpha + \sum_{i=1}^n y_i}{(\beta + n)^2} \right)^{1/2} \approx 0.060.$$

eBay data - Prior and Posterior of θ



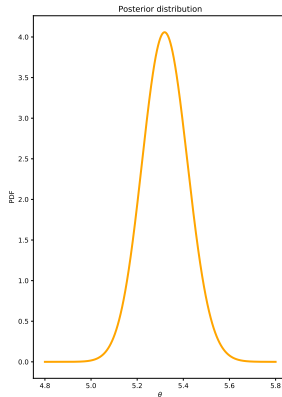
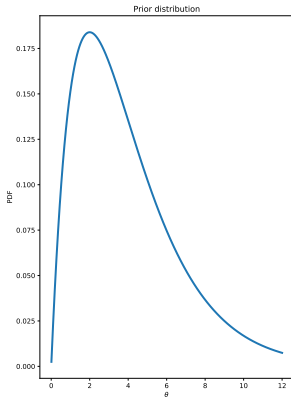
eBay data - Fit



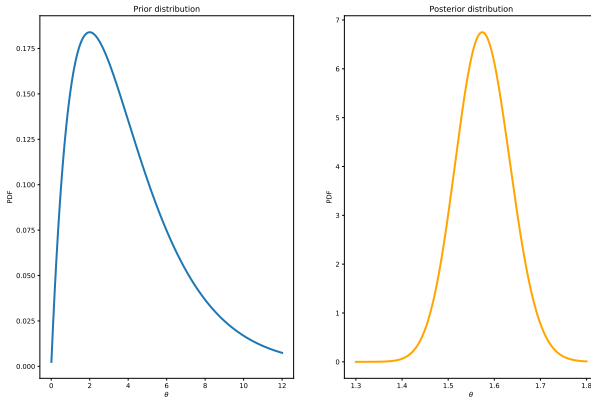
eBay - low/high seller's reservation price

- The data is very heterogenous. Some auctions start with very high reservations prices (lowest price accepted by the seller).
- Split the data into auctions with low/high reservation prices.
- **Low reservation price auctions:**
 - ▶ $n = 550$ eBay coin auctions.
 - ▶ Posterior mean: 5.321 bids.
- **High reservation price auctions:**
 - ▶ $n = 450$ eBay coin auctions.
 - ▶ Posterior mean: 1.576 bids.

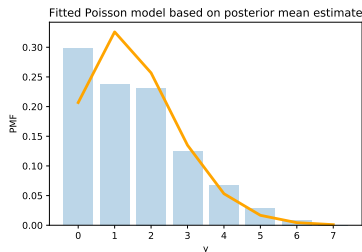
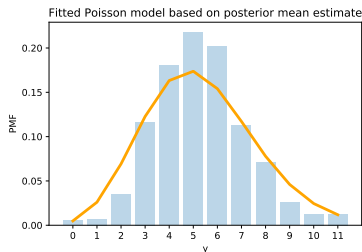
eBay - low seller's reservation price



eBay - high seller's reservation price



eBay - fit low/high reservation prices



- Better fits, but still not good enough.
- Lab 3: Fit **Poisson regression** with reservation price as continuous covariate.

Posterior intervals

- **Bayesian 95% credible interval**: the probability that the unknown parameter θ lies in the interval is 0.95.

- Approximate 95% **credible interval** for θ

$$E(\theta|y) \pm 1.96 \cdot SD(\theta|y) = [3.517; 3.753]$$

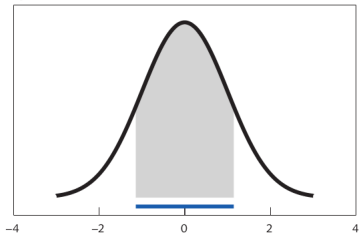
- An exact 95% **equal-tail interval** is $[3.518; 3.754]$

- **Highest Posterior Density (HPD)** interval contains the θ values with highest pdf.

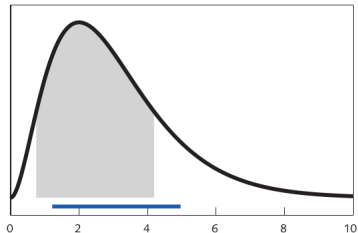
$$[3.518; 3.752]$$

Illustration of different interval types

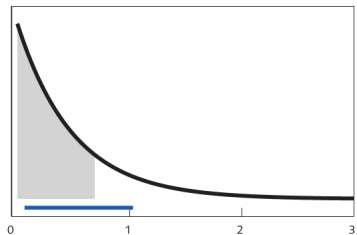
Symmetrical distribution



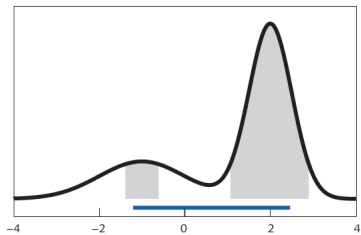
Skewed distribution



Skewed monotonous distribution



Bimodal distribution



Conjugate priors

- Normal likelihood: Normal prior \rightarrow Normal posterior.
- Bernoulli likelihood: Beta prior \rightarrow Beta posterior.
- Poisson likelihood: Gamma prior \rightarrow Gamma posterior.
- **Conjugate priors**: A prior is conjugate to a model if the prior and posterior belong to the same distributional family.
- Formal definition: Let $\mathcal{F} = \{p(y|\theta), \theta \in \Theta\}$ be a class of sampling distributions. A family of distributions \mathcal{P} is **conjugate** for \mathcal{F} if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}$$

holds for all $p(y|\theta) \in \mathcal{F}$.

Prior elicitation

- 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 elicited prior to the expert.
- Beware of **psychological effects**, such as anchoring.

Prior elicitation - AR(p) example

- Autoregressive process of order p

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

- 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 $\phi_i \sim N(\mu_i, \psi_i)$, $i = 1, \dots, p$ are independent a priori.
- Prior on $\phi = (\phi_1, \dots, \phi_p)$ centered on persistent AR(1) process: $\mu_1 = 0.8, \mu_2 = \dots = \mu_p = 0$
- $\text{Var}(\phi_i) = \frac{c}{i^\lambda}$. “Longer” lags are more likely to be zero a priori.

Jeffreys' prior

- **Fisher information** (the amount of information that $\mathbf{x} = (x_1, \dots, x_n)$ carries about θ):

$$I(\theta) = -E_{\mathbf{x}|\theta} \left(\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2} \right)$$

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

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

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

Jeffreys' prior for Bernoulli sampling

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

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

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

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

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

Thus, the Jeffreys' prior is

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

Jeffreys' prior for negative binomial sampling

- Jeffreys' prior:

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

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

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

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

- Thus, the Jeffreys' prior is

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

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

Different types of prior information

- Real **expert information**. Combo of previous studies and experience.
- **Vague prior** information.
- **Smoothness priors**. Regularization. Shrinkage. Big thing in modern statistics/machine learning.