Bayesian Learning - Lecture 1

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Course overview

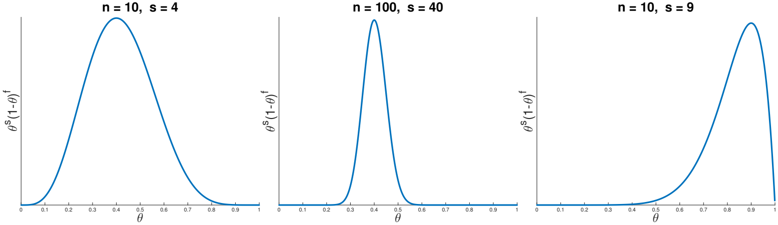
- Three audiences:
  - **Master students in Statistics and Data Mining** (732A91)
  - **Engineering students** (TDDE07)
  - **PhD students**
- Course **webpage** is **here**. Course **syllabus** is **here**.
- Modes of teaching:
  - **Lectures** (Mattias Villani)
  - **Mathematical exercises** (Mattias Villani)
  - **Computer labs** (Måns Magnusson)
- **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, 3 credits
  - Computer exam (using R), 3 credits

Lecture overview

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

THE 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.
$$p(x_1, ..., x_n|\theta) = p(x_1|\theta) \cdots p(x_n|\theta) = \theta^s (1-\theta)^f$$
- **Maximum likelihood estimator** $\hat{\theta}$ maximizes $p(x_1, ..., x_n|\theta)$.
- Given the data $x_1, ..., x_n$, we may plot $p(x_1, ..., x_n|\theta)$ as a function of θ.



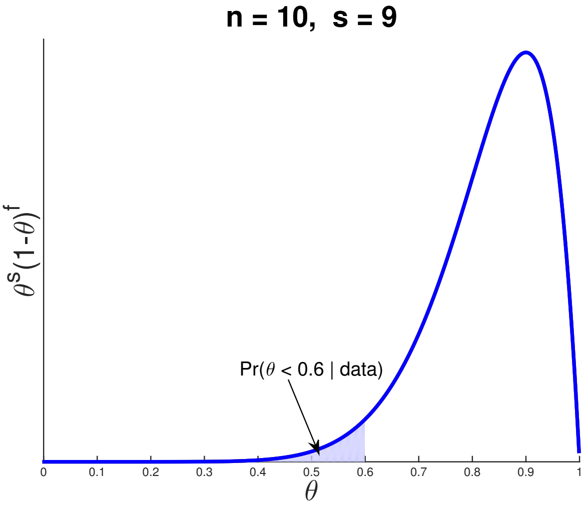
The likelihood function

- Say it out loud:

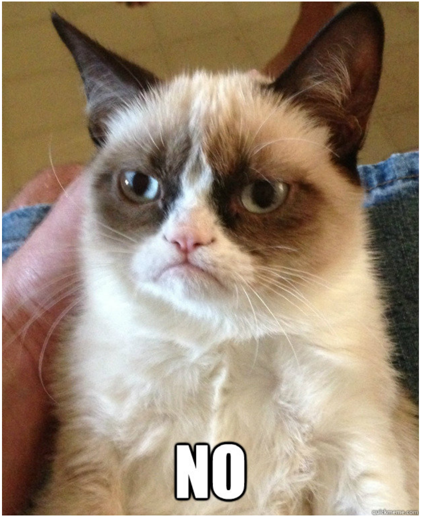
  The likelihood function is
  the probability of the observed data
  considered as a function of the parameter.

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

Probabilities from the likelihood!!

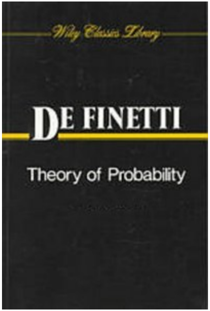


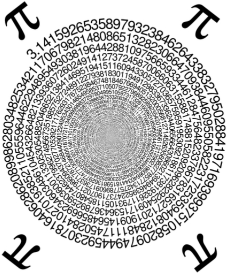
PROBABILITIES FROM THE LIKELIHOOD!!



Uncertainty and subjective probability

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





Bayesian learning

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

Learning from data - Bayes' theorem

- How do we **update** from the **prior** $p(\theta)$ to the **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

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

- The prior $p(\theta)$ is the hero that converts the likelihood function $p(Data|\theta)$ into a posterior probability density $p(\theta|Data)$.
- A probability distribution for θ is extremely useful. **Decision making**.
- **No prior - no posterior - no useful inferences - no fun**.

Bayes' theorem for medical diagnosis

- A $=$ {Horrible and 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 given a positive test:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A^c)p(A^c)} \approx 0.001797.$$

- Probably not sick, but 18 times more probable than before the test.

- Morale of the story: 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 that makes $p(\theta|Data)$ integrate 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

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

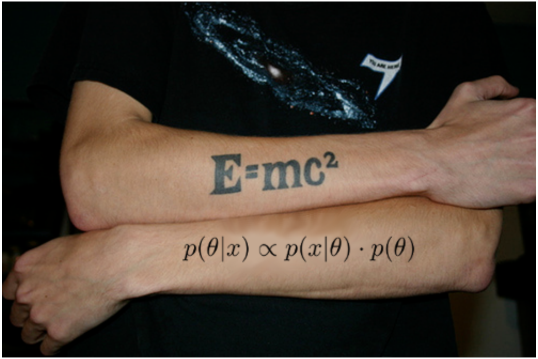
Great theorems make great tattoos

- All you need to know:

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

or

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



Bernoulli trials - Beta prior

- **Model**

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

- **Prior**

$$\theta \sim Beta(\alpha, \beta)$$

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

- **Posterior**

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

- This is proportional to the $Beta(\alpha+s, \beta+f)$ density.
- The **prior-to-posterior** mapping reads

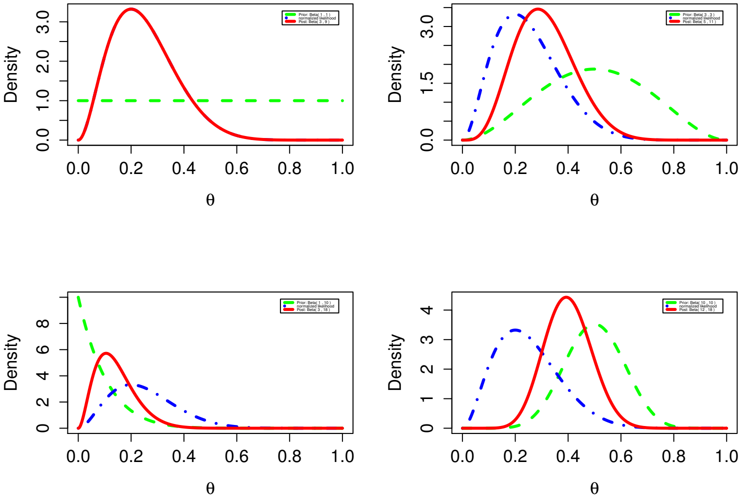
$$\theta \sim Beta(\alpha, \beta) \overset{x_1, ..., x_n}{\Longrightarrow} \theta | x_1, ..., x_n \sim Beta(\alpha+s, \beta+f).$$

Bernoulli example: spam emails

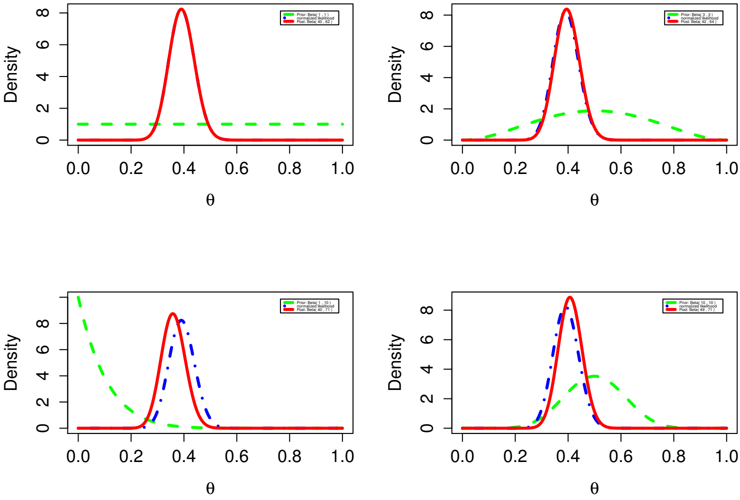
- George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam.
- Let $x_i = 1$ if i:th email is spam. Assume $x_i|\theta \overset{iid}{\sim} Bernoulli(\theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$.
- Posterior

$$\theta|x \sim Beta(\alpha + 1813, \beta + 2788)$$

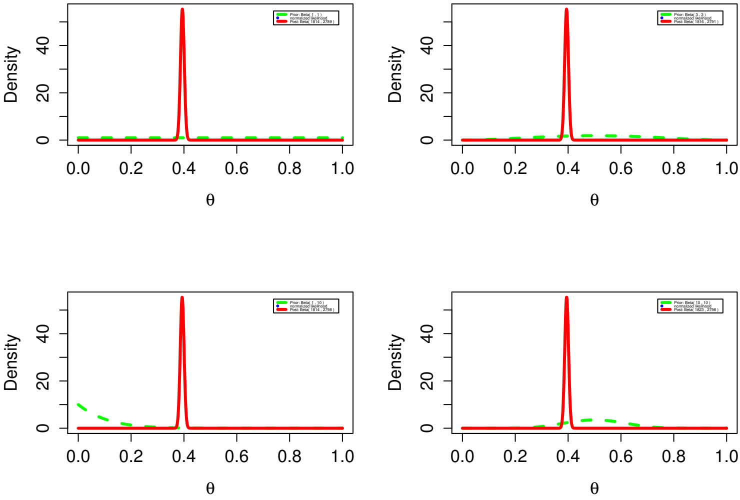
SPAM DATA (N=10): PRIOR SENSITIVITY



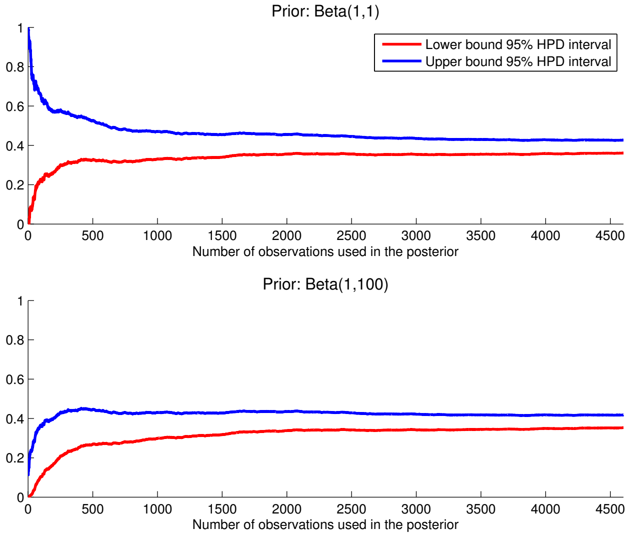
Spam data (n=100): prior sensitivity



Spam data (n=4601): prior sensitivity



Spam data: Posterior convergence



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

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

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

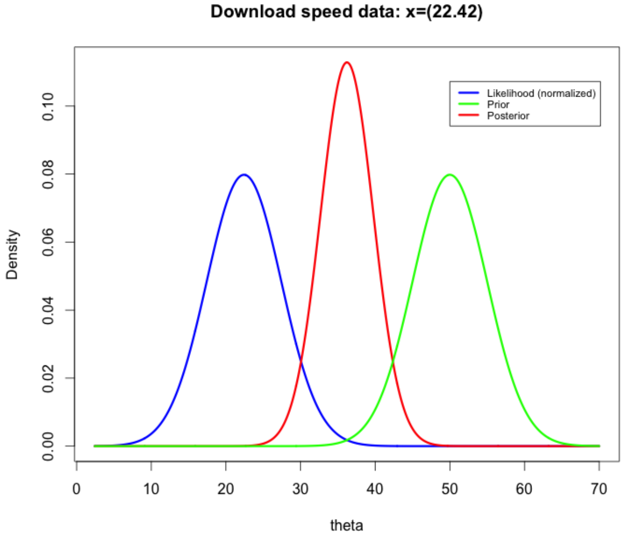
$$\text{Posterior mean} =$$

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

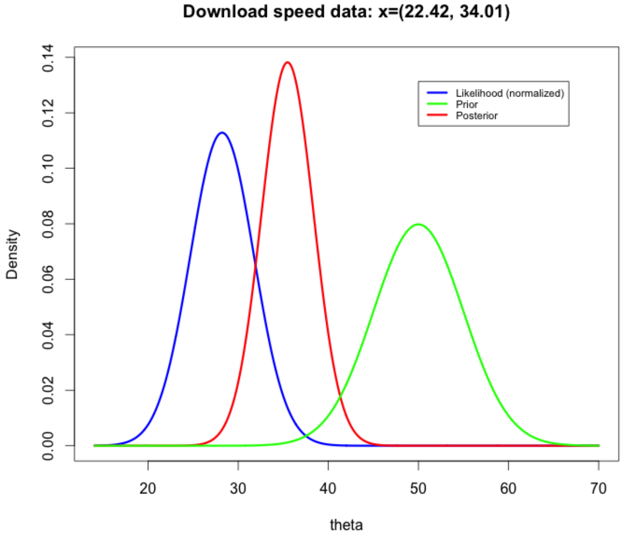
Download speed

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

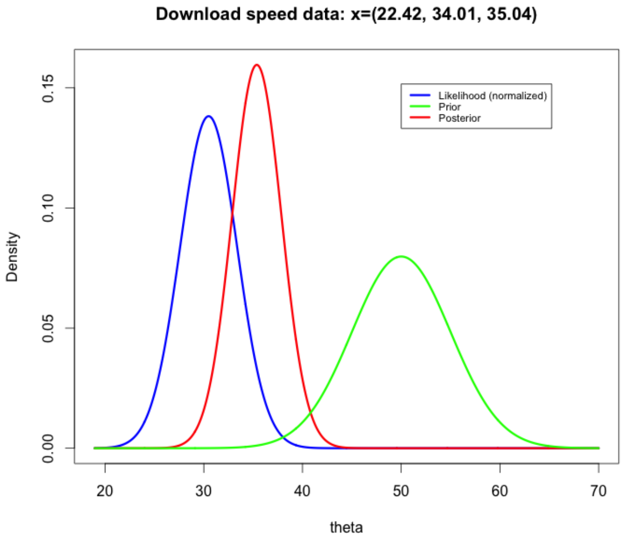
Download speed n=1



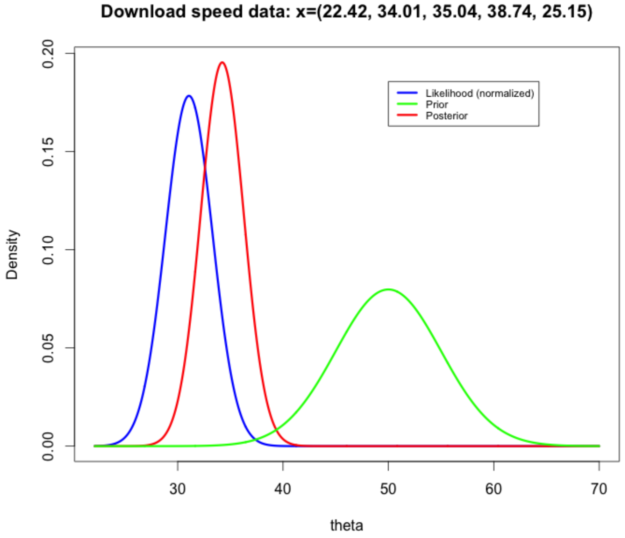
Download speed n=2



DOWNLOAD SPEED N=3

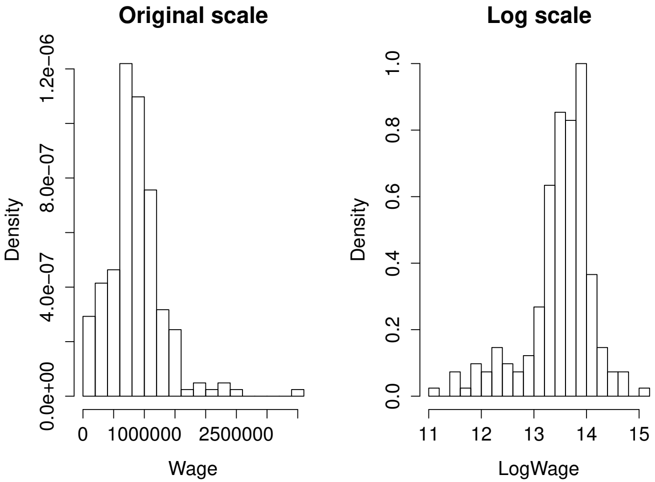


Download speed n=5



Canadian wages data

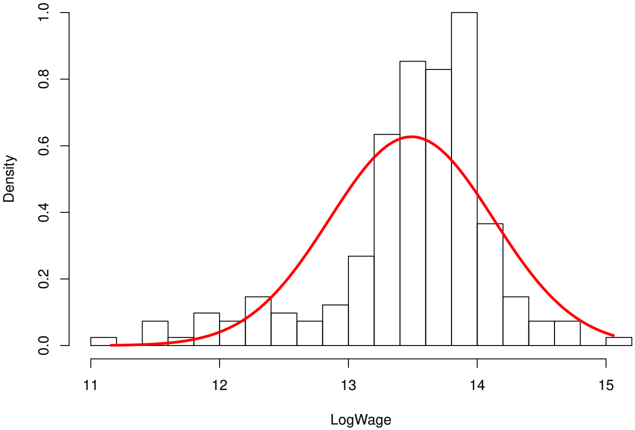
- Data on wages for 205 Canadian workers.



Canadian wages

- Model
$$X_1, ..., X_n|\theta \sim N(\theta, \sigma^2), \; \sigma^2 = 0.4$$
- Prior
$$\theta \sim N(\mu_0, \tau_0^2), \; \mu_0 = 12 \text{ and } \tau_0 = 10$$
- Posterior
$$\theta|x_1, ..., x_n \sim N\left(\mu_n, \tau_n^2\right),$$
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



Code to play with

- tt
- fff