

# Bayesian statistics, simulation and software

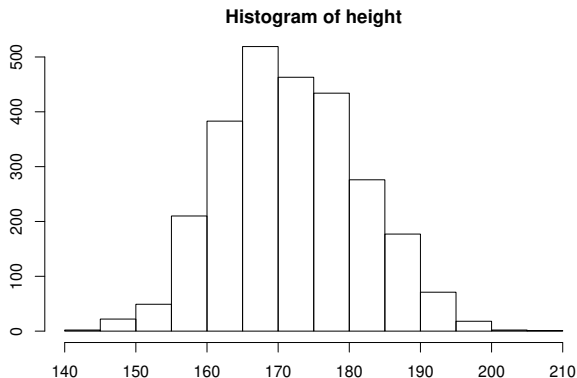
## Module 4: Normal model, improper and conjugate priors

Jesper Møller and Ege Rubak

Department of Mathematical Sciences  
Aalborg University

# Another example: normal sample with known precision

Heights of some Copenhageners in 1995:

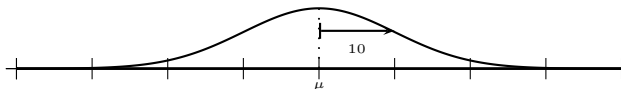


**Assume:** Heights are independent and normal:  $X_i \sim \mathcal{N}(\mu, \tau)$ ,  $i = 1, \dots, n$ .

**For now:** Assume precision  $\tau$  is known.

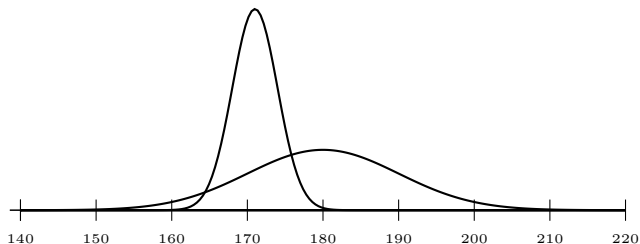
# Bayesian idea: Illustration

**Data model:**  $X \sim \mathcal{N}(\mu, 0.01)$  (i.e. pop. sd = 10)



**Prior:** We believe that the population mean is most likely between 160 cm and 200 cm:  $\pi(\mu) \sim \mathcal{N}(180, 0.01)$  (i.e.  $160 \leq \mu \leq 200$  with prior probability 95%).

**Posterior:** After observing a number of heights, our knowledge about  $\mu$  is updated/summarised by the posterior (is it mostly about females?):



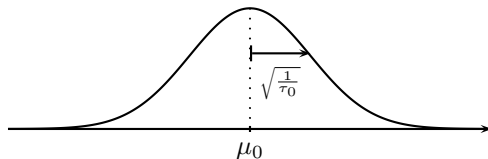
# Normal example: One (!) observation

Data model:  $X \sim \mathcal{N}(\mu, \tau)$

Assume: Precision  $\tau$  known.

Interest: The unknown mean  $\mu$ .

**Prior:** The prior for  $\mu$  is specified as  $\mu \sim \mathcal{N}(\mu_0, \tau_0)$ .



(Within 1 sd =  $\sqrt{1/\tau_0}$  from  $\mu_0$  with prior probability 68%.)

## Normal example: Data density

**Data:** One observation,  $X = x$ , from  $\mathcal{N}(\mu, \tau)$ :

$$\begin{aligned}\pi(x|\mu) &= \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x - \mu)^2\right) \\ &= \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau x^2 - \frac{1}{2}\tau\mu^2 + \tau\mu x\right) \\ &\propto \exp\left(-\frac{1}{2}\tau x^2 + \tau\mu x\right)\end{aligned}$$

where "proportional to" ( $\propto$ ) refers to that we consider  $\mu$  as fixed whilst  $x$  is the argument for this conditional density.

**Notice** the "pattern" inside the last exponential. If we instead pay attention to the likelihood, i.e. when we consider  $x$  as fixed, we get

$$L(\mu; x) \propto \exp\left(-\frac{1}{2}\tau\mu^2 + \tau\mu x\right).$$

## Normal example: Posterior density

*Posterior*  $\propto$  *Likelihood*  $\times$  *Prior*:

$$\begin{aligned}\pi(\mu|x) &\propto L(\mu; x)\pi(\mu) \\ &= L(\mu; x)\sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{1}{2}\tau_0(\mu - \mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\tau\mu^2 + \tau x\mu - \frac{1}{2}\tau_0\mu^2 + \tau_0\mu\mu_0\right) \\ &= \exp\left(-\frac{1}{2}(\tau + \tau_0)\mu^2 + (\tau x + \tau_0\mu_0)\mu\right) \\ &\sim \mathcal{N}\left(\frac{\tau x + \tau_0\mu_0}{\tau + \tau_0}, \tau + \tau_0\right).\end{aligned}$$

**Notice:** Both prior and posterior for  $\mu$  are normal (conjugateness), and posterior precision = prior precision + likelihood precision.

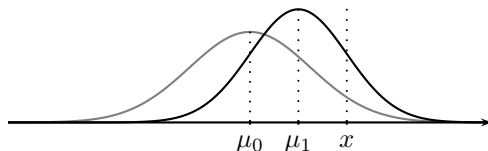
# Normal example: Posterior mean & variance

The posterior:  $\pi(\mu|x) \sim \mathcal{N}\left(\frac{\tau x + \tau_0 \mu_0}{\tau + \tau_0}, \tau + \tau_0\right)$ .

**Posterior expectation:**

$$\mu_1 \equiv \mathbb{E}[\mu|x] = \frac{\tau x + \tau_0 \mu_0}{\tau + \tau_0} = \frac{\tau}{\tau + \tau_0} x + \frac{\tau_0}{\tau + \tau_0} \mu_0.$$

*Weighted average* of prior mean and observation  $x$ .



**Posterior variance** is smaller than prior variance:

$$\frac{1}{\tau_1} \equiv \text{Var}[\mu|x] = \frac{1}{\tau + \tau_0} < \frac{1}{\tau_0} = \text{Var}(\mu).$$

# Posterior as prior — or updating believes

**General setup:** We are interested in parameter  $\theta$ .

- Data model:  $\pi(x|\theta)$ .
- Prior:  $\pi(\theta)$ .
- Data: First observation  $x_1 \sim \pi(x_1|\theta)$ .
- Posterior:  $\pi(\theta|x_1) \propto \pi(x_1|\theta)\pi(\theta)$ .

Assume we have a second observation  $x_2 \sim \pi(x_2|\theta)$  which is (conditional) independent of  $x_1$  given  $\theta$ . Then the **new posterior** is

$$\begin{aligned}\pi(\theta|x_1, x_2) &\propto \pi(x_1, x_2|\theta)\pi(\theta) \\ &= \pi(x_1|\theta)\pi(x_2|\theta)\pi(\theta) \\ &\propto \underbrace{\pi(x_2|\theta)}_{\text{likelihood}} \underbrace{\pi(\theta|x_1)}_{\text{new prior}}\end{aligned}$$

**Notice:** The posterior after observing  $x_1$  is the prior before observing  $x_2$ .



# Independent normal case

Posterior mean and precision after one observation  $x_1$ :

$$\mu_1 = \frac{x_1\tau + \mu_0\tau_0}{\tau + \tau_0} \quad \text{and} \quad \tau_1 = \tau + \tau_0.$$

Next,  $\mu_1$  and  $\tau_1$  are prior mean and precision before observing  $x_2$ .

Then posterior mean and precision after observing (conditionally independent)  $x_1$  and  $x_2$  are

$$\begin{aligned}\mu_2 &= \mathbb{E}[\mu|x_1, x_2] = \frac{x_2\tau + \mu_1\tau_1}{\tau + \tau_1} \\ &= \frac{x_2\tau + x_1\tau + \mu_0\tau_0}{\tau + \tau + \tau_0} = \frac{(x_1 + x_2)\tau + \mu_0\tau_0}{2\tau + \tau_0}, \\ \tau_2 &= \text{Var}[\mu|x_1, x_2] = \tau + \tau_1 = 2\tau + \tau_0.\end{aligned}$$

This can easily be generalised...

# Many independent normal observations

Assume:

- $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau)$ .
- $\tau > 0$  is known.
- Prior  $\pi(\mu) \sim \mathcal{N}(\mu_0, \tau_0)$ .

The posterior is then

$$\pi(\mu|x_1, x_2, \dots, x_n) \sim \mathcal{N}\left(\frac{\tau \sum_i x_i + \tau_0 \mu_0}{n\tau + \tau_0}, n\tau + \tau_0\right).$$

# Posterior mean: Sanity check

The posterior is

$$\pi(\mu|x_1, x_2, \dots, x_n) \sim \mathcal{N}\left(\frac{\tau \sum_{i=1}^n x_i + \tau_0 \mu_0}{n\tau + \tau_0}, n\tau + \tau_0\right).$$

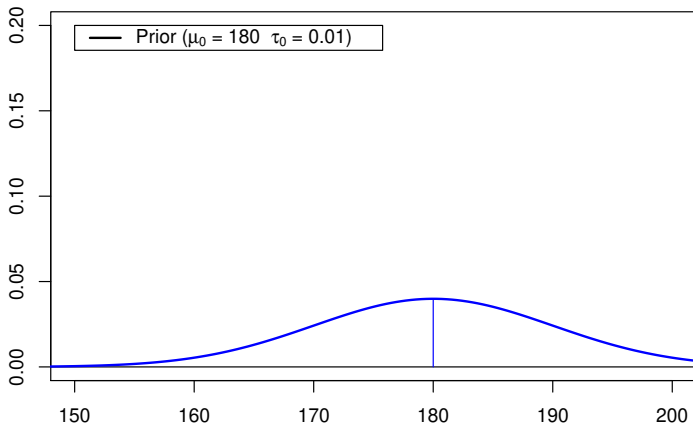
Does the posterior mean make sense? Yes, because

$$\begin{aligned}\mu_n := \mathbb{E}[\mu|x_1, \dots, x_n] &= \frac{\tau \sum_{i=1}^n x_i + \tau_0 \mu_0}{n\tau + \tau_0} \\ &= \frac{\tau n \frac{1}{n} \sum_{i=1}^n x_i + \tau_0 \mu_0}{n\tau + \tau_0} \\ &= \frac{n\tau}{n\tau + \tau_0} \bar{x} + \frac{\tau_0}{n\tau + \tau_0} \mu_0\end{aligned}$$

is a weighted average of sample average  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and prior mean  $\mu_0$ , so that for  $n$  large,  $\mu_n \approx \bar{x}$ , and so the choice of  $\mu_0$  and  $\tau_0$  is of little importance.

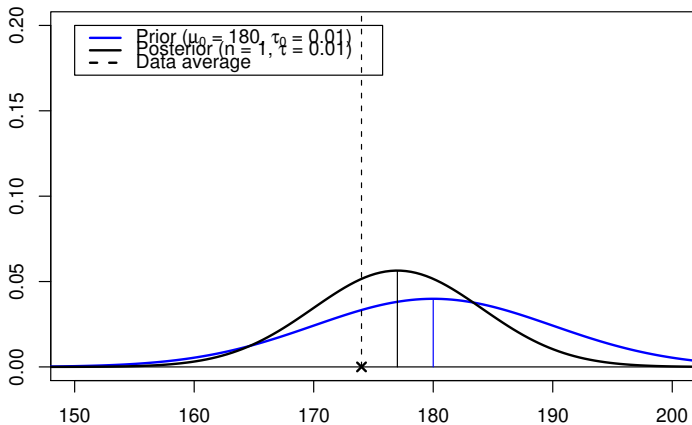
NB: precision/knowledge  $\tau_n = n\tau + \tau_0$  is ever more precise as  $n$  increases.

# Heights in Copenhagen: Prior



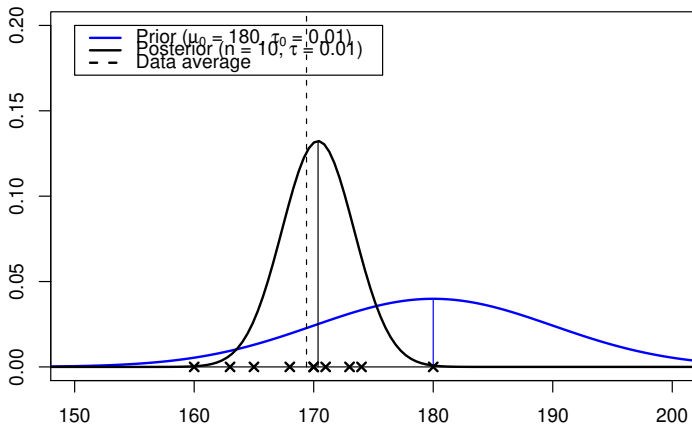
# Heights in Copenhagen: Posterior

One observation



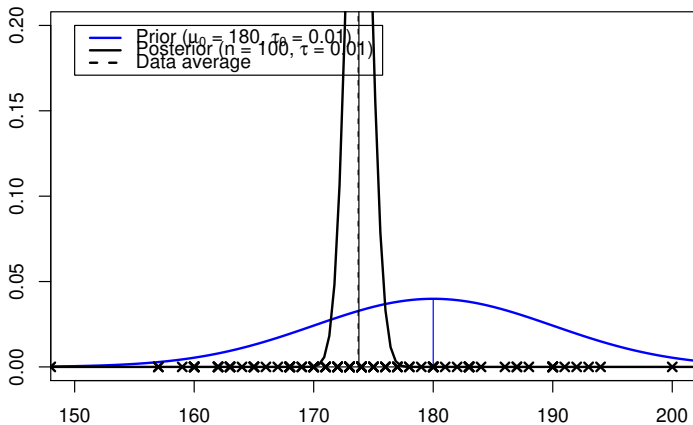
# Heights in Copenhagen: Posterior

Ten observations



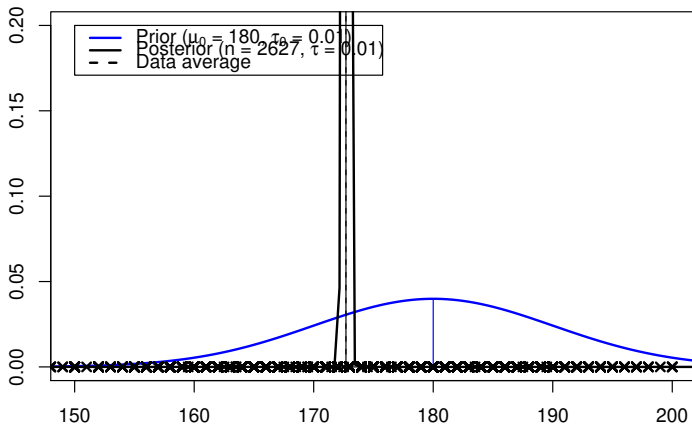
# Heights in Copenhagen: Posterior

100 observations



# Heights in Copenhagen: Posterior

2627 observations





# How to summarise the posterior $\pi(\theta|x)$ ?

The posterior is usually summarised using one or more of the following:

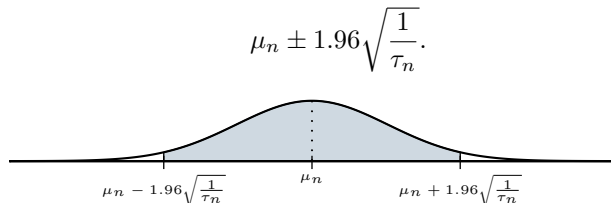
- Plot of posterior density  $\pi(\theta|x)$ . See previous slides.
- Posterior mean and variance/precision.
- Credible intervals. See next slide.
- Maximum a posteriori (MAP) estimate

$$MAP(\theta) = \underset{\theta}{\operatorname{argmax}} \pi(\theta|x).$$

# Credible intervals

Methods for defining a suitable (e.g.) 95% credible interval/region for a one-dimensional (sub-)parameter  $\theta$  include:

- Highest posterior density region (HPDI): Choosing the narrowest region(!) which contains  $\theta$  with 95% posterior probability.
- Central posterior interval (CPI): The interval given by the 2.5% and 97.5% quantiles of the posterior distribution for  $\theta$ .
- In case of the normal example,  $\mu|x_1, \dots, x_n \sim \mathcal{N}(\mu_n, \tau_n)$ , so the 95% HPDI (= CPI) for  $\mu$  is



## Compared to classical confidence interval

The classical 95% confidence interval for  $\mu$  is

$$\bar{x} \pm 1.96 \sqrt{\frac{1}{n\tau}}.$$

For CPI: Assuming the prior precision is  $\tau_0 = 0$ , i.e. an infinite prior variance (an improper prior, but the posterior is well-defined/limiting case), then

$$\mu_n = \bar{x}, \quad \tau_n = n\tau,$$

i.e. the same as the classical confidence interval  
— but different interpretations!!

# Conjugate priors

In this example: Both prior and posterior were normal distributions! This is very convenient — and we say that the prior and posterior distributions are *conjugate* distributions.

## Definition: Conjugate priors

Let  $\pi(x|\theta)$  be the data model. A class  $\Pi$  of prior distributions for  $\theta$  is said to be **conjugate** for  $\pi(x|\theta)$  if

$$\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta) \in \Pi$$

whenever  $\pi(\theta) \in \Pi$ . That is, the prior and posterior are in the same class of distributions.

**Notice:**  $\Pi$  should be a class of “tractable” distributions for this to be useful.

# Improper prior

If we have no prior knowledge, we may (perhaps) be tempted to use a “flat” prior, i.e.

$$\pi(\theta) \propto k \quad (\text{a positive constant}).$$

If  $\theta \in \mathbf{R}$ , this is an example of an improper prior, because

$$\int_{-\infty}^{\infty} \pi(\theta) d\theta \propto \int_{-\infty}^{\infty} k = \infty.$$

Perhaps problematic but may not be considered to be an issue if posterior is proper, i.e. if

$$\int \pi(\theta|x) d\theta \propto \int \pi(x|\theta)\pi(\theta) d\theta < \infty.$$

**Notice:** If  $\pi(\theta) \propto 1$ , then

MAP estimator = maximum likelihood estimator.

## Normal example: Unknown precision, known mean

- **Data model:**  $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau)$ :

$$\pi(x|\tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\tau \sum_{i=1}^n (x_i - \mu)^2\right).$$

- **Prior:** Gamma distribution:  $\pi(\tau) \sim \text{Gamma}(\alpha, \beta)$ , that is

$$\pi(\tau) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \tau^{\alpha-1} \exp\left(-\frac{\tau}{\beta}\right), \quad \tau > 0.$$

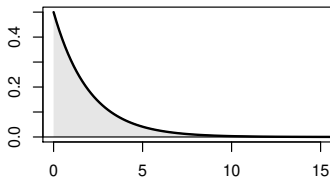
Properties:

$$\mathbb{E}[\tau] = \alpha\beta, \quad \text{Var}[\tau] = \alpha\beta^2.$$

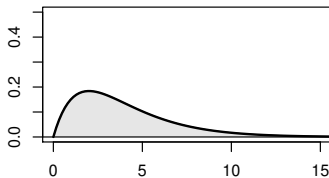
Shape parameter  $\alpha > 0$  and scale parameter  $\beta > 0$  (see next slide).

# Gamma distribution

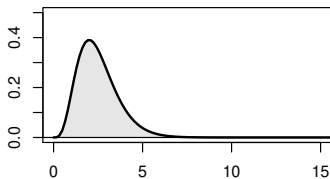
$\alpha=1, \beta=2$



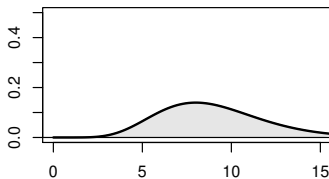
$\alpha=2, \beta=2$



$\alpha=5, \beta=0.5$



$\alpha=9, \beta=1$



## Normal example: Posterior precision

- **Data model:**  $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau)$ :
- **Prior:**  $\pi(\tau) \sim \text{Gamma}(\alpha, \beta)$ .
- **Posterior:** An easy calculation gives

$$\pi(\tau|x) \sim \text{Gamma} \left( \frac{n}{2} + \alpha, \left\{ \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 + \frac{1}{\beta} \right\}^{-1} \right).$$

Posterior mean and variance:

$$\mathbb{E}[\tau|x] = \frac{\frac{n}{2} + \alpha}{\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 + \frac{1}{\beta}}, \quad \text{Var}[\tau|x] = \frac{\frac{n}{2} + \alpha}{\left( \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 + \frac{1}{\beta} \right)^2}.$$

For large  $n$ ,

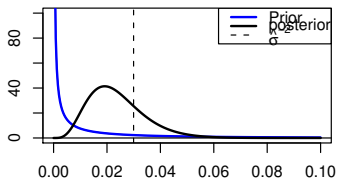
$$\mathbb{E}[\tau|x] \approx \frac{1}{\hat{\sigma}^2} \quad \text{where } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

is the maximum likelihood estimate of the variance.

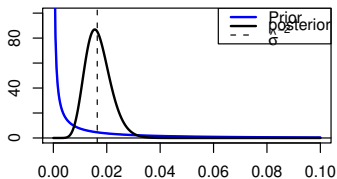


# Known mean: Priors and posteriors

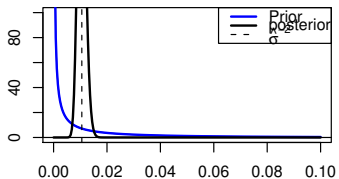
$\alpha=0.1, \beta=0.1, n=10$



$\alpha=0.1, \beta=0.1, n=25$



$\alpha=0.1, \beta=0.1, n=100$



$\alpha=0.1, \beta=0.1, n=2627$

