

# Bayesian statistics, simulation and software

## Module 10: Bayesian prediction and model checking

Jesper Møller and Ege Rubak

Department of Mathematical Sciences  
Aalborg University

# Prior predictions

Suppose we want to predict future data  $\tilde{x}$  *without* observing any data  $x$ .

**Assume:**

- **Data model:**  $\tilde{x}|\theta \sim x|\theta \sim \pi(x|\theta)$ .
- **Prior:**  $\pi(\theta)$ .

This implies a joint distribution:

$$(\tilde{x}, \theta) \sim \pi(x, \theta) = \pi(x|\theta)\pi(\theta).$$

From this joint distribution we obtain the marginal density of  $\tilde{x}$ ,

$$\tilde{x} \sim \pi(x) = \int \pi(x|\theta)\pi(\theta)d\theta,$$

which is called the **prior predictive density**.

## Prior prediction: Normal case, $\tau$ known

### Assume:

- **Data model:**  $\pi(x|\mu) \sim \mathcal{N}(\mu, \tau)$ .
- **Prior:**  $\pi(\mu) \sim \mathcal{N}(\mu_0, \tau_0)$ .

Prior predictive density:

$$\begin{aligned}\pi(x) &= \int \pi(x|\mu)\pi(\mu)d\mu \\ &= \int \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right) \sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{1}{2}\tau_0(\mu-\mu_0)^2\right) d\mu\end{aligned}$$

and a simple calculation shows

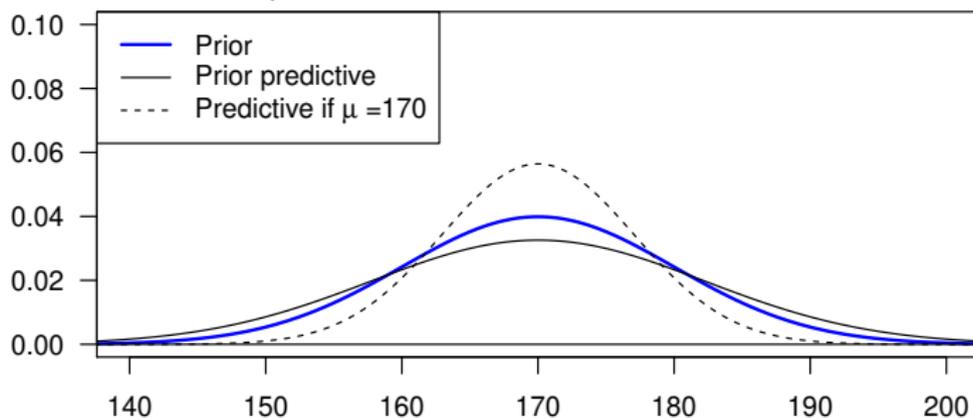
$$\pi(x) \propto \exp\left(-\frac{1}{2} \frac{\tau\tau_0}{\tau+\tau_0} (x-\mu_0)^2\right) \sim \mathcal{N}\left(\mu_0, \frac{\tau\tau_0}{\tau+\tau_0}\right).$$

Easier argument:  $x - \mu \sim \mathcal{N}(0, \tau)$  is independent of  $\mu \sim \mathcal{N}(\mu_0, \tau_0)$ , so  $x \sim \mathcal{N}\left(\mu_0, \left(\frac{1}{\tau} + \frac{1}{\tau_0}\right)^{-1}\right) = \mathcal{N}\left(\mu_0, \frac{\tau\tau_0}{\tau+\tau_0}\right)$ .

NB: prior predictive precision is smaller than that for prior and for data.

# Prior predictive distribution

Illustration of the fact that prior predictive precision  $<$  prior precision (ignore the dashed line):



# Simulating the prior predictive distribution

If the prior predictive density  $\pi(x)$  is difficult to derive we can simply make a simulation  $\tilde{x}$  in two steps:

1. Generate parameter from prior:  $\theta \sim \pi(\theta)$ .
2. Conditional on  $\theta$  generate  $\tilde{x}$ :  $\tilde{x} \sim \pi(x|\theta)$ .

# Posterior prediction: one observation

Now, suppose we have observed data  $x$  and want to predict a possible *future* observation  $\tilde{x}$  given data  $x$ .

**Assume:**

- **Data model:**  $\tilde{x}|\theta \sim x|\theta \sim \pi(x|\theta)$ , and given  $\theta$  then  $x$  and  $\tilde{x}$  are independent.
- **Prior:**  $\pi(\theta)$ .

The joint density of predicted data  $\tilde{x}$ , data  $x$  and parameter  $\theta$  is

$$\begin{aligned}\pi(\tilde{x}, x, \theta) &= \pi(\tilde{x}, x | \theta)\pi(\theta) = \pi(\tilde{x}|\theta)\pi(x|\theta)\pi(\theta) \\ &= \pi(\tilde{x}|\theta)\pi(\theta|x)\pi(x)\end{aligned}$$

where  $\pi(\tilde{x}|\theta)$  and  $\pi(x|\theta)$  represent the same conditional distribution (namely that given by the data model).

**Definition:** The **posterior predictive distribution** is the distribution of  $\tilde{x}$  conditional on data  $x$ :

$$\pi(\tilde{x}|x) = \int \pi(\tilde{x}, \theta|x)d\theta = \int \frac{\pi(\tilde{x}, \theta, x)}{\pi(x)}d\theta = \int \pi(\tilde{x}|\theta)\pi(\theta|x)d\theta.$$

Thus we have now replaced the prior density with the posterior density.

# Simulating the posterior predictive distribution based on one observation

If the posterior predictive density  $\pi(\tilde{x}|x)$  is difficult to derive we can simply make a simulation  $\tilde{x}$  in two steps:

1. Generate parameter from posterior:  $\theta|x \sim \pi(\theta|x)$ .
2. Conditional on  $\theta$  generate  $\tilde{x}$  from data model:  $\tilde{x} \sim \pi(x|\theta)$ .

# Posterior prediction: iid observations

Suppose (given  $\theta$ ) we have iid data  $x_1, \dots, x_n$  and want to predict a possible *future* observation  $\tilde{x}$  given data  $\mathbf{x} = (x_1, \dots, x_n)$ .

**Assume:**

- Given  $\theta$ , then  $x_1, \dots, x_n, x$  are iid.
- Prior:  $\pi(\theta)$ .

The joint density of predicted data  $\tilde{x}$ , data  $\mathbf{x}$  and parameter  $\theta$  is

$$\begin{aligned}\pi(\tilde{x}, \mathbf{x}, \theta) &= \pi(\theta)\pi(\tilde{x}|\theta) \prod_{i=1}^n \pi(x_i|\theta) = \pi(\tilde{x}|\theta)\pi(\theta)\pi(\mathbf{x}|\theta) \\ &= \pi(\tilde{x}|\theta)\pi(\theta|\mathbf{x})\pi(\mathbf{x}).\end{aligned}$$

**Definition:** The **posterior predictive distribution** is the distribution of  $\tilde{x}$  conditional on data  $\mathbf{x}$ :

$$\pi(\tilde{x}|\mathbf{x}) = \int \pi(\tilde{x}, \theta|\mathbf{x})d\theta = \int \frac{\pi(\tilde{x}, \theta, \mathbf{x})}{\pi(\mathbf{x})}d\theta = \int \pi(\tilde{x}|\theta)\pi(\theta|\mathbf{x})d\theta.$$

Thus we have again replaced the prior density with the posterior density.

# Simulating the posterior predictive distribution based on iid observations

If the posterior predictive density  $\pi(\tilde{x}|\mathbf{x})$  is difficult to derive we can simply make a simulation  $\tilde{x}$  in two steps:

1. Generate parameter from posterior:  $\theta|\mathbf{x} \sim \pi(\theta|\mathbf{x})$ .
2. Conditional on  $\theta$  generate  $\tilde{x}$  from the data model for one observation  $x_i$  ( $1 \leq i \leq n$ ):  $\tilde{x} \sim \pi(x_i|\theta)$ .

## Posterior prediction: Normal case, $\tau$ known

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

**Prior:**  $\pi(\mu) \sim \mathcal{N}(\mu_0, \tau_0)$ .

**Posterior:**  $\pi(\mu|\mathbf{x}) \sim \mathcal{N}(\mu_1, \tau_1)$ ,  $\mu_1 = \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}$  and  $\tau_1 = n\tau + \tau_0$ .

Conditioned on  $\mathbf{x}$ , we have that  $\tilde{x} - \mu \sim \mathcal{N}(0, \tau)$  is independent of  $\mu$ , so  $\tilde{x} = (\tilde{x} - \mu) + \mu$  satisfies

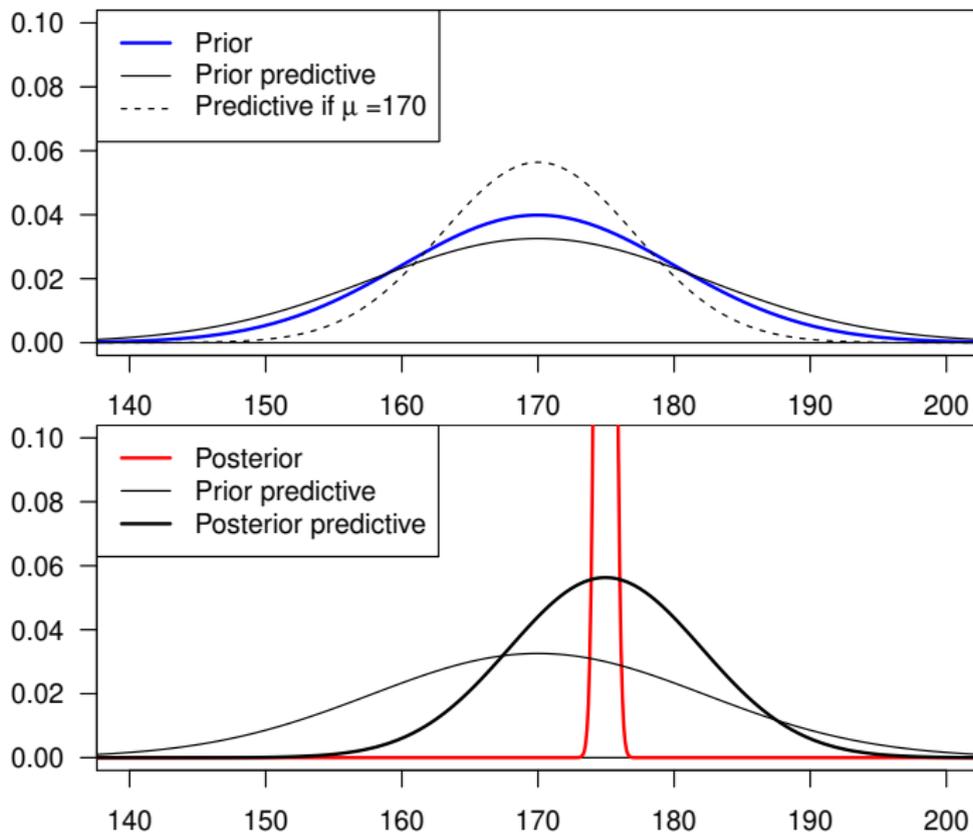
$$\tilde{x}|\mathbf{x} \sim \mathcal{N}\left(\mu_1, (\tau^{-1} + \tau_1^{-1})^{-1}\right) = \mathcal{N}\left(\mu_1, \frac{\tau\tau_1}{\tau + \tau_1}\right).$$

NB: Posterior predictive mean and posterior mean are equal. Posterior predictive precision  $\frac{\tau\tau_1}{\tau + \tau_1}$  is like the prior predictive precision but with  $\tau_0$  replaced by  $\tau_1$ , and it is smaller than posterior precision  $\tau_1$  and than data precision  $\tau$ , but larger than prior predictive precision  $\frac{\tau\tau_0}{\tau + \tau_0}$ .

When  $n$  is large, we have  $\tilde{x}|\mathbf{x} \stackrel{approx}{\sim} \mathcal{N}(\bar{x}, \tau)$ .

When  $\tau_0 = 0$  (i.e. we consider an improper prior), we have (as when making predictions in classical statistics)  $\tilde{x}|\mathbf{x} \stackrel{approx}{\sim} \mathcal{N}(\bar{x}, \tau)$  (for  $n$  large).

# Prior and posterior predictive distributions



# Model checking

**Idea:** If the model is correct, then posterior predictions of the data should look like the observed data. **Difficulty:** How to choose a good measure of “similarity”?

**Example:** We have observed a sequence of  $n = 20$  zeros and ones:

1 1 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0

**Model:**  $X_1, X_2, \dots, X_{20}$  are IID where  $P(X_i = 1) = p$  is unknown.

**Prior:**  $\pi(p) \sim Be(\alpha, \beta)$  where  $\alpha > 0$  and  $\beta > 0$  are known.

**Posterior:**  $\pi(p|\mathbf{x}) \sim Be(\#\text{ones} + \alpha, \#\text{zeros} + \beta)$ .

**Model checking:** We simulate  $N$  posterior predictive realisations

$$\tilde{\mathbf{X}}^{(i)} = (\tilde{X}_1^{(i)}, \tilde{X}_2^{(i)}, \dots, \tilde{X}_{20}^{(i)}) \quad i = 1, \dots, N.$$

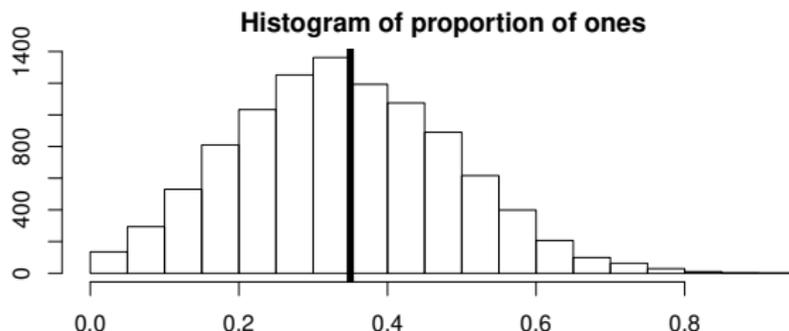
If these vectors look “similar” to the data above, it would indicate that the model is probably okay.

# Model checking: First attempt (a failure)

Define summary function

$$s(\mathbf{x}) = \#\text{ones in } \mathbf{x}.$$

Histogram for  $s(\tilde{\mathbf{x}}^{(i)})$  for  $N = 10,000$  independent posterior predictions:



So the observed number of ones is in no way unusual compared to the posterior predictions.

This is just as expected — so we need another summary function  $s(\mathbf{x})$ .

# Model checking: Second attempt (a success)

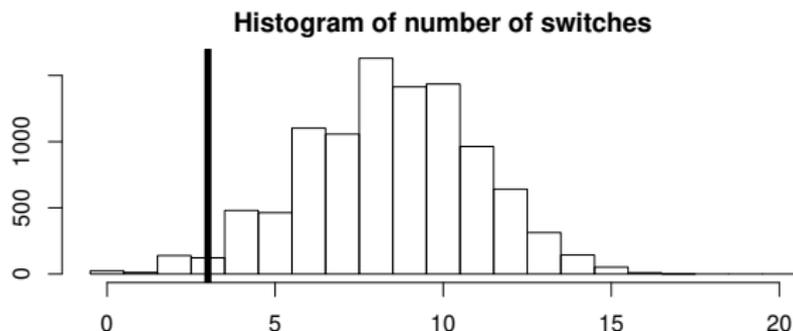
Define summary function

$s(\mathbf{x}) =$  number of switches between ones and zeros in  $\mathbf{x}$ .

In the data the number of switches is 3:

1 1 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0

Histogram for  $s(\tilde{\mathbf{x}}^{(i)})$  for  $N = 10,000$  independent posterior predictions:

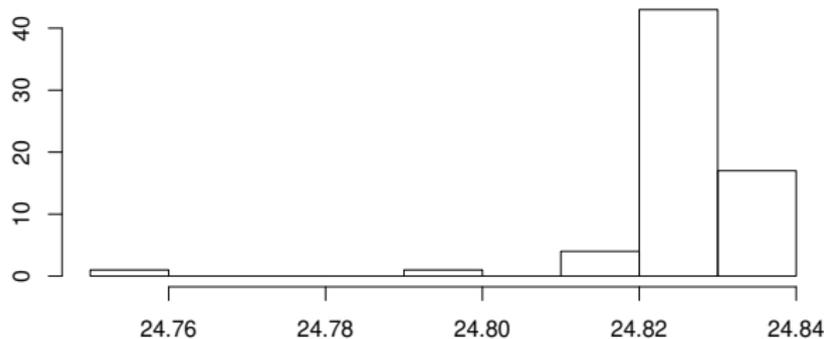


Only around 1.7% of the posterior prediction have 3 or fewer switches.

This suggests that the model assumption of independence is questionable.

## Example: Speed of light

66 measurements of the time it takes light to travel 7445 meters  
(deviations in nanoseconds from a given number):



Data model:

$$x_1, \dots, x_{66} \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau).$$

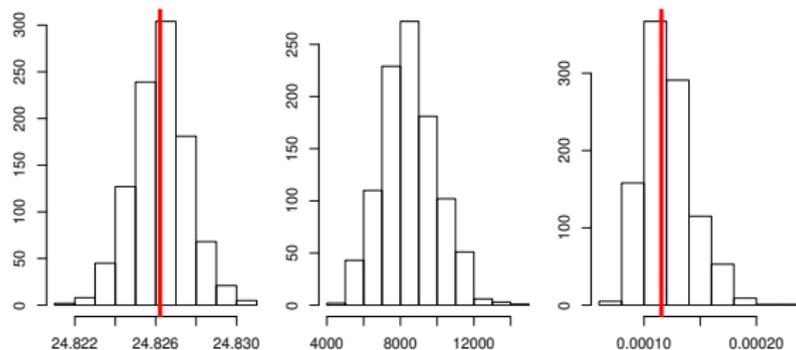
(Questionable?)

Prior:

$$\pi(\mu, \tau) \sim \mathcal{N}(0, 0.001) \times \text{Gamma}(0.001, 1000).$$

# Example: Speed of light

Posterior distribution of  $\mu$ ,  $\tau$  and  $1/\tau$ :



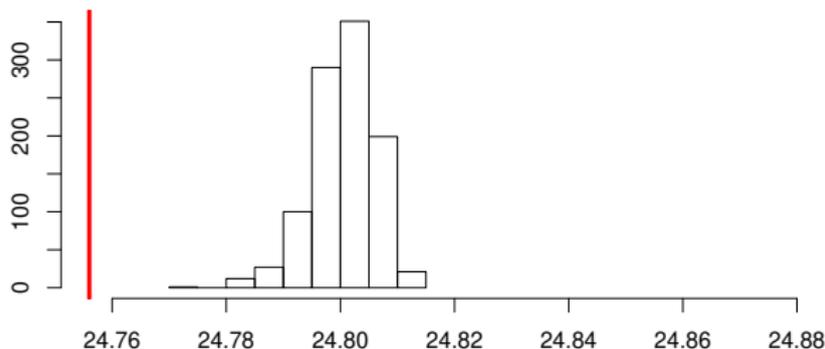
Red lines denote sample mean and sample variance, respectively.

## Example: Speed of light

Data contain one very low measurement. Is this unusual?

Generate 1000 posterior predictive samples  $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_{66}^{(i)})$ ,  $i = 1, \dots, 1000$ , and define

$$s(\mathbf{x}) = \min\{x_1, \dots, x_{66}\}.$$



Conclusion: The smallest value in the data is very unlikely under the assumed model.