



Newcastle
University

MAS3902

Bayesian Inference

Semester 2, 2022–23

Dr. Lee Fawcett

School of Mathematics, Statistics & Physics

Course overview

You were introduced to the Bayesian approach to statistical inference in MAS2903. This module showed statistical analysis in a very different light to the frequentist approach used in other courses. The frequentist approach bases inference on the sampling distribution of (usually unbiased) estimators; as you may recall, the Bayesian framework combines information expressed as expert subjective opinion with experimental data. You have probably realised that the Bayesian approach has many advantages over the frequentist approach. In particular it provides a more natural way of dealing with parameter uncertainty and inference is far more straightforward to interpret.

Much of the work in this module will be concerned with extending the ideas presented in MAS2903 to more realistic models with many parameters that you may encounter in real life situations. These notes are split into four chapters:

- **Chapter 1** reviews some of the key results for Bayesian inference of single parameter problems studied in Stage 2.
- **Chapter 2** studies the case of a random sample from a normal population and determines how to make inferences about the population mean and precision, and about future values from the population. The Group Project is based on this material.
- **Chapter 3** contains some general results for multi-parameter problems. You will encounter familiar concepts, such as how to represent *vague prior information* and the *asymptotic normal posterior distribution*.
- **Chapter 4** introduces *Markov chain Monte Carlo* techniques which have truly revolutionised the use of Bayesian inference in applications. Inference proceeds by simulating realisations from the posterior distribution. The ideas will be demonstrated using an R library specially written for the module. This material is extended in the 4th year Bayesian Statistics module.

Contents

1	Single parameter problems	1
1.1	Prior and posterior distributions	1
1.2	Different levels of prior knowledge	10
1.2.1	Substantial prior knowledge	10
1.2.2	Limited prior knowledge	10
1.2.3	Vague prior knowledge/prior ignorance	10
1.3	The asymptotic posterior distribution	14
1.4	Bayesian inference	15
1.4.1	Estimation	15
1.4.2	Prediction	19
1.5	Sequential Bayesian updating	23
1.6	Learning objectives	27
2	Inference for a normal population	29
2.1	Bayes Theorem for many parameters	29
2.2	Prior to posterior analysis	31
2.2.1	Marginal distributions	32
2.3	Confidence intervals and regions	37
2.4	Predictive distribution	41
2.5	Summary	44
2.6	Why do we have so many different distributions?	45
2.7	Learning objectives	46
3	General results for multi-parameter problems	47
3.1	Different levels of prior knowledge	47
3.2	Asymptotic posterior distribution	52

3.3	Learning objectives	55
4	Non-conjugate multi-parameter problems	57
4.1	Why is inference not straightforward in non-conjugate problems?	57
4.2	Simulation-based inference	59
4.3	Motivation for MCMC methods	62
4.4	The Gibbs sampler	62
4.4.1	Processing output from a Gibbs sampler	63
4.4.2	Bayesian inference using a Gibbs sampler	69
4.5	Metropolis-Hastings sampling	81
4.5.1	Symmetric chains (Metropolis method)	82
4.5.2	Independence chains	86
4.6	Hybrid methods	86
4.6.1	Componentwise transitions	86
4.6.2	Metropolis within Gibbs	87
4.7	Summary	92
4.8	Learning objectives	93
A	Summary of distributions	95
A.1	Distributions for data	95
A.2	Distributions for prior beliefs	99
	Exercises	101
	The R package <code>nclbayes</code>	107

Chapter 1

Single parameter problems

This chapter reviews some of the key results for Bayesian inference of single parameter problems studied in MAS2903.

1.1 Prior and posterior distributions

Suppose we have data $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ which we model using the probability (density) function $f(\mathbf{x}|\theta)$, which depends on a single parameter θ . Once we have observed the data, $f(\mathbf{x}|\theta)$ is the *likelihood function* for θ and is a function of θ (for fixed $\mathbf{x}$) rather than of $\mathbf{x}$ (for fixed θ).

Also, suppose we have prior beliefs about likely values of θ expressed by a probability (density) function $\pi(\theta)$. We can combine both pieces of information using the following version of Bayes Theorem. The resulting distribution for θ is called the posterior distribution for θ as it expresses our beliefs about θ *after* seeing the data. It summarises all our current knowledge about the parameter θ .

Using Bayes Theorem, the posterior probability (density) function for θ is

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x})}$$

where

$$f(\mathbf{x}) = \begin{cases} \int_{\Theta} \pi(\theta) f(\mathbf{x}|\theta) d\theta & \text{if } \theta \text{ is continuous,} \\ \sum_{\Theta} \pi(\theta) f(\mathbf{x}|\theta) & \text{if } \theta \text{ is discrete.} \end{cases}$$

Also, as $f(\mathbf{x})$ is not a function of θ , Bayes Theorem can be rewritten as

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta) \times f(\mathbf{x}|\theta)$$

i.e. Posterior $\propto$ Prior $\times$ Likelihood.

Example 1.1

Table 1.1 shows some data on the number of cases of foodborne botulism in England and Wales. It is believed that cases occur at random at a constant rate θ in time (a Poisson process) and so can be modelled as a random sample from a Poisson distribution with mean θ .

Year	1998	1999	2000	2001	2002	2003	2004	2005
Cases	2	0	0	0	1	0	2	1

Table 1.1: Number of cases of foodborne botulism in England and Wales, 1998–2005

An expert in the epidemiology of similar diseases was consulted to construct a suitable prior distribution for the rate θ .

- (a) What might be an appropriate probability model for the rate θ ? Why is this termed the *prior distribution*?



A gamma distribution might be appropriate here as θ is a rate and must be positive; the gamma distribution has support over the positive real line. This is termed the *prior distribution* as it represents our beliefs about θ before observing any data.

- (b) Suppose the epidemiologist tells you that she expects around two cases of foodborne botulism per year in England and Wales. After further discussions with the epidemiologist, you are also able to elicit that $SD(\theta) \approx \sqrt{2}$. Using your answer to part (a) find your full prior distribution for θ , and write down its density.



Suppose our prior is $\theta \sim Ga(g, h)$. Then from the information given, we can use the formulae for the gamma mean and standard deviation to write down:

$$E(\theta) = \frac{g}{h} = 2 \quad (\text{using the expert's most likelt value for } \theta)$$

and

$$SD(\theta) = \frac{\sqrt{g}}{h} = \sqrt{2}.$$

Solving simulataneously for g and h gives $g = 2$ and $h = 1$, and so we have $\theta \sim Ga(2, 1)$. Therefore:

$$\pi(\theta) = \frac{1^2 \theta^{2-1} e^{-1\theta}}{\Gamma(2)} = \propto \theta e^{-\theta}, \quad \theta > 0.$$

(c) Determine the posterior distribution for θ .



We know that

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}.$$

We know from the last question that the prior density is $\pi(\theta) \propto \theta e^{-\theta}$. What about the Poisson likelihood? From the data provided, we can write down that

$$\begin{aligned} f(\mathbf{x}|\theta) &= \frac{e^{-\theta}\theta^2}{2!} \times \frac{e^{-\theta}\theta^0}{0!} \times \cdot \times \frac{e^{-\theta}\theta^1}{1!} \\ &\propto e^{-8\theta}\theta^6. \end{aligned}$$

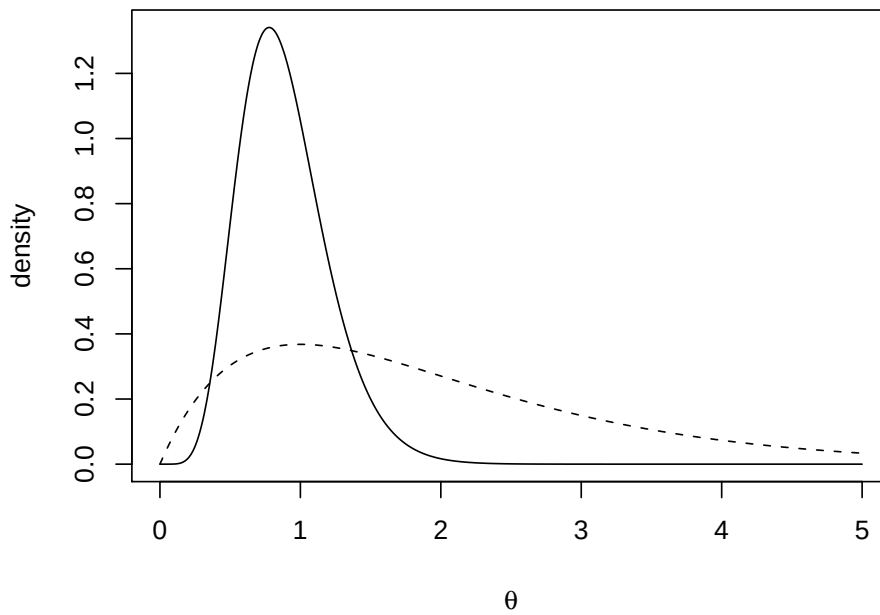
Therefore

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto \pi(\theta) \times f(\mathbf{x}|\theta) \\ &\propto \theta e^{-\theta} \times \theta^6 e^{-8\theta} \\ &\propto \theta^7 e^{-9\theta} \quad \theta > 0 \end{aligned}$$

and we recognise this as being proportional to the density of a $Ga(8, 9)$ distribution; that is, $\theta|\mathbf{x} \sim Ga(8, 9)$.

Thus the data have updated our beliefs about θ from a $Ga(2, 1)$ distribution to a $Ga(8, 9)$ distribution. Plots of these distributions are given in Figure 1.1, and Table 1.2 gives a summary of the main changes induced by incorporating the data — a $Ga(g, h)$ distribution has mean g/h , variance g/h^2 and mode $(g - 1)/h$.

Notice that, as the mode of the likelihood function is close to that of the prior distribution, the information in the data is consistent with that in the prior distribution. Also there is a reduction in variability from the prior to the posterior distributions. The similarity between the prior beliefs and the data has reduced the uncertainty we have about the rate θ at which cases occur.

Figure 1.1: Prior (dashed) and posterior (solid) densities for θ

	Prior	Likelihood	Posterior
$Mode(\theta)$	1.00	0.75	0.78
$E(\theta)$	2.00	—	0.89
$SD(\theta)$	1.41	—	0.31

Table 1.2: Changes in beliefs about θ

Example 1.2

Consider now the general case of Example 1.1: suppose $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \dots, n$ (independent) and our prior beliefs about θ are summarised by a $Ga(g, h)$ distribution (with g and h known), with density

$$\pi(\theta) = \frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \quad (1.1)$$

Determine the posterior distribution for θ .





Solution to Example 1.2 continued...

The likelihood function for θ is

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!}, \quad \theta > 0 \\ &\propto \theta^{n\bar{x}} e^{-n\theta}, \quad \theta > 0. \end{aligned} \quad (1.2)$$

Using Bayes Theorem, the posterior density function is

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto \pi(\theta) f(\mathbf{x}|\theta) \\ &\propto \frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \theta^{n\bar{x}} e^{-n\theta}, \quad \theta > 0 \\ \text{i.e. } \pi(\theta|\mathbf{x}) &= k \theta^{g+n\bar{x}-1} e^{-(h+n)\theta}, \quad \theta > 0 \end{aligned} \quad (1.3)$$

where k is a constant that does not depend on θ . Therefore, the posterior density takes the form $k\theta^{G-1}e^{-H\theta}$, $\theta > 0$ and so the posterior must be a gamma distribution. Thus we have $\theta|\mathbf{x} \sim Ga(G = g + n\bar{x}, H = h + n)$.

Summary:

If we have a random sample from a $Po(\theta)$ distribution and our prior beliefs about θ follow a $Ga(g, h)$ distribution then, after incorporating the data, our (posterior) beliefs about θ follow a $Ga(g + n\bar{x}, h + n)$ distribution.

The changes in our beliefs about θ are summarised in Table 1.3, taking $g \geq 1$. Notice

	Prior (1.1)	Likelihood (1.2)	Posterior (1.3)
$Mode(\theta)$	$(g - 1)/h$	$\bar{x}$	$(g + n\bar{x} - 1)/(h + n)$
$E(\theta)$	g/h	–	$(g + n\bar{x})/(h + n)$
$SD(\theta)$	$\sqrt{g}/h$	–	$\sqrt{g + n\bar{x}}/(h + n)$

Table 1.3: Changes in beliefs about θ

that the posterior mean is greater than the prior mean if and only if the likelihood mode is greater than the prior mean, that is,

$$E(\theta|\mathbf{x}) > E(\theta) \iff Mode_{\theta}\{f(\mathbf{x}|\theta)\} > E(\theta).$$

The standard deviation of the posterior distribution is smaller than that of the prior distribution if and only if the sample mean is not too large, that is

$$SD(\theta|\mathbf{x}) < SD(\theta) \iff Mode_{\theta}\{f(\mathbf{x}|\theta)\} < \left(2 + \frac{n}{h}\right) E(\theta),$$

and this will be true in large samples.

Example 1.3

Suppose we have a random sample from a normal distribution. In Bayesian statistics, when dealing with the normal distribution, the mathematics is more straightforward working with the precision ($= 1/\text{variance}$) of the distribution rather than the variance itself. So we will assume that this population has unknown mean μ but known precision τ : $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \dots, n$ (independent), where τ is known. Suppose our prior beliefs about μ can be summarised by a $N(b, 1/d)$ distribution, with probability density function

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}. \quad (1.4)$$

Determine the posterior distribution for μ .

Hint:

$$d(\mu - b)^2 + n\tau(\bar{x} - \mu)^2 = (d + n\tau) \left\{ \mu - \left(\frac{db + n\tau\bar{x}}{d + n\tau} \right) \right\}^2 + c$$

where c does not depend on μ .



The likelihood function for μ is

$$\begin{aligned} f(\mathbf{x}|\mu) &= \prod_{i=1}^n \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x_i - \mu)^2\right\} \\ &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2\right\}. \end{aligned}$$

Now

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_{i=1}^n \{(x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2(x_i - \bar{x})(\bar{x} - \mu)\} \\ &= \sum_{i=1}^n \{(x_i - \bar{x})^2 + (\bar{x} - \mu)^2\} + 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \bar{x}) \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \end{aligned}$$

Let $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ and so

$$\sum_{i=1}^n (x_i - \mu)^2 = n[s^2 + (\bar{x} - \mu)^2].$$



Solution to Example 1.3 continued...

Therefore

$$f(\mathbf{x}|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2} [s^2 + (\bar{x} - \mu)^2]\right\}. \quad (1.5)$$

Using Bayes Theorem, the posterior density function is, for $\mu \in \mathbb{R}$

$$\begin{aligned} \pi(\mu|\mathbf{x}) &\propto \pi(\mu) f(\mathbf{x}|\mu) \\ &\propto \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \\ &\quad \times \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2} [s^2 + (\bar{x} - \mu)^2]\right\} \\ &\propto \exp\left\{-\frac{1}{2} [d(\mu - b)^2 + n\tau(\bar{x} - \mu)^2]\right\} \\ &\propto \exp\left\{-\frac{1}{2} \left[(d + n\tau) \left\{\mu - \left(\frac{db + n\tau\bar{x}}{d + n\tau}\right)\right\}^2 + c\right]\right\} \end{aligned}$$

using the hint

$$\propto \exp\left\{-\frac{1}{2} \left[(d + n\tau) \left\{\mu - \left(\frac{db + n\tau\bar{x}}{d + n\tau}\right)\right\}^2\right]\right\}$$

as c does not depend on μ . Let

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \quad \text{and} \quad D = d + n\tau. \quad (1.6)$$

Then

$$\pi(\mu|\mathbf{x}) = k \exp\left\{-\frac{D}{2}(\mu - B)^2\right\}, \quad (1.7)$$

where k is a constant that does not depend on μ . Therefore, the posterior density takes the form $k \exp\{-D(\mu - B)^2/2\}$, $\mu \in \mathbb{R}$ and so the posterior distribution must be a normal distribution: we have $\mu|\mathbf{x} \sim N(B, 1/D)$.

Summary:

If we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known) and our prior beliefs about μ follow a $N(b, 1/d)$ distribution then, after incorporating the data, our (posterior) beliefs about μ follow a $N(B, 1/D)$ distribution.

The changes in our beliefs about μ are summarised in Table 1.4. Notice that the posterior

	Prior (1.4)	Likelihood (1.5)	Posterior (1.7)
$Mode(\mu)$	b	$\bar{x}$	$(db + n\tau\bar{x})/(d + n\tau)$
$E(\mu)$	b	–	$(db + n\tau\bar{x})/(d + n\tau)$
$Precision(\mu)$	d	–	$d + n\tau$

Table 1.4: Changes in beliefs about μ

mean is greater than the prior mean if and only if the likelihood mode (sample mean) is greater than the prior mean, that is

$$E(\mu|\mathbf{x}) > E(\mu) \iff Mode_{\mu}\{f(\mathbf{x}|\mu)\} > E(\mu).$$

Also, the standard deviation of the posterior distribution is smaller than that of the prior distribution.

Example 1.4

The 18th century physicist Henry Cavendish made 23 experimental determinations of the earth's density, and these data (in g/cm^3) are given below.

5.36	5.29	5.58	5.65	5.57	5.53	5.62	5.29
5.44	5.34	5.79	5.10	5.27	5.39	5.42	5.47
5.63	5.34	5.46	5.30	5.78	5.68	5.85	

Suppose that Cavendish asserts that the error standard deviation of these measurements is $0.2 g/cm^3$, and assume that they are normally distributed with mean equal to the true earth density μ . Using a normal prior distribution for μ with mean $5.41 g/cm^3$ and standard deviation $0.4 g/cm^3$, derive the posterior distribution for μ .



From the data we calculate $\bar{x} = 5.4848$ and $s = 0.1882$. Therefore, the assumed standard deviation $\sigma = 0.2$ is probably okay. We also have $\tau = 1/0.2^2$, $b = 5.41$, $d = 1/0.4^2$ and $n = 23$.



Solution to Example 1.4 continued...

Therefore, using Example 1.3, the posterior distribution is $\mu|\mathbf{x} \sim N(B, 1/D)$, where

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} = \frac{5.41/0.4^2 + 23 \times 5.4848/0.2^2}{1/0.4^2 + 23/0.2^2} = 5.4840$$

and

$$D = d + n\tau = \frac{1}{0.4^2} + \frac{23}{0.2^2} = \frac{1}{0.0415^2}.$$

Therefore the posterior distribution is $\mu|\mathbf{x} \sim N(5.484, 0.0415^2)$ and is shown in Figure 1.2.

The actual mean density of the earth is 5.515 g/cm^3 (Wikipedia). We can determine the (posterior) probability that the mean density is within 0.1 of this value as follows. The posterior distribution is $\mu|\mathbf{x} \sim N(5.484, 0.0415^2)$ and so

$$Pr(5.415 < \mu < 5.515|\mathbf{x}) = 0.9510,$$

calculated using the R command `pnorm(5.515, 5.484, 0.0415) - pnorm(5.415, 5.484, 0.0415)`.

Without the data, the only basis for determining the earth's density is via the prior distribution. Here the prior distribution is $\mu \sim N(5.4, 0.4^2)$ and so the (prior) probability that the mean density is within 0.2 of the (now known) true value is

$$Pr(5.315 < \mu < 5.515) = 0.1896,$$

calculated using the R command `pnorm(5.515, 5.4, 0.4) - pnorm(5.315, 5.4, 0.4)`.

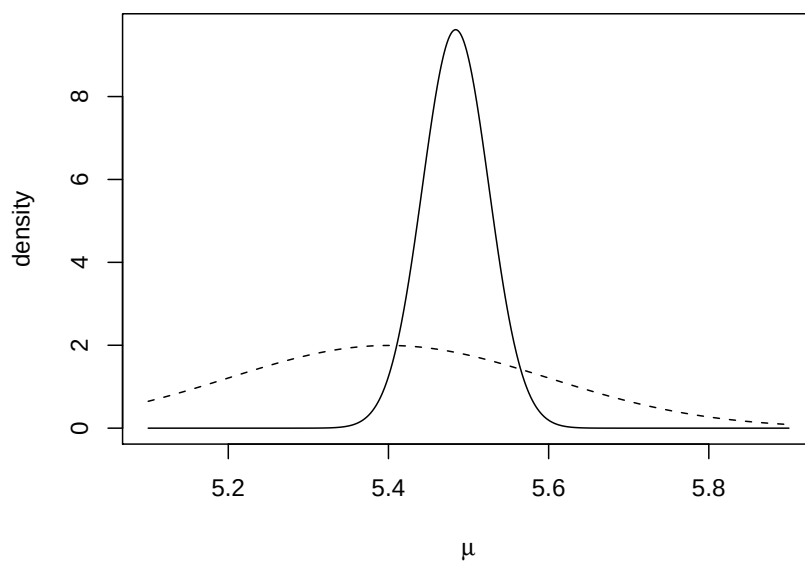


Figure 1.2: Prior (dashed) and posterior (solid) densities for the earth's density

1.2 Different levels of prior knowledge

1.2.1 Substantial prior knowledge

We have *substantial prior information* for θ when the prior distribution *dominates* the posterior distribution, that is $\pi(\theta|\mathbf{x}) \sim \pi(\theta)$.

When we have substantial prior information there can be some difficulties:

1. the intractability of the mathematics in deriving the posterior distribution — though with modern computing facilities this is less of a problem,
2. the practical formulation of the prior distribution — coherently specifying prior beliefs in the form of a probability distribution is far from straightforward.

1.2.2 Limited prior knowledge

When prior information about θ is limited, the pragmatic approach is to choose a distribution which makes the Bayes updating from prior to posterior mathematically straightforward, and use what prior information is available to determine the parameters of this distribution. For example

- Poisson random sample, Gamma prior distribution $\longrightarrow$ Gamma posterior distribution
- Normal random sample (known variance), Normal prior distribution $\longrightarrow$ Normal posterior distribution

In these examples, the prior distribution and the posterior distribution come from the same family. This leads us to the following definition.

Definition 1.1

Suppose that data $\mathbf{x}$ are to be observed with distribution $f(\mathbf{x}|\theta)$. A family $\mathfrak{F}$ of prior distributions for θ is said to be *conjugate* to $f(\mathbf{x}|\theta)$ if for every prior distribution $\pi(\theta) \in \mathfrak{F}$, the posterior distribution $\pi(\theta|\mathbf{x})$ is also in $\mathfrak{F}$.

Notice that the conjugate family depends crucially on the model chosen for the data $\mathbf{x}$. For example, the only family conjugate to the model “random sample from a Poisson distribution” is the Gamma family.

1.2.3 Vague prior knowledge/prior ignorance

If we have very little or no prior information about the model parameters θ , we must still choose a prior distribution in order to operate Bayes Theorem. Obviously, it would

be sensible to choose a prior distribution which is not concentrated about any particular value, that is, one with a very large variance. In particular, most of the information about θ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\mathbf{x}) \sim f(\mathbf{x}|\theta)$.

We often represent vague prior knowledge by using a prior distribution which is conjugate to the model for $\mathbf{x}$ and which is as diffuse as possible, that is, has as large a variance as possible.

We may attempt to represent prior ignorance using a *Jeffreys prior*, specified in terms of Fisher's Information:

$$I(\theta) = E_{\mathbf{X}|\theta} \left[-\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{X}|\theta) \right]. \quad (1.8)$$

Specifically, Jeffreys recommended that we represent prior ignorance by the prior distribution

$$\pi(\theta) \propto \sqrt{I(\theta)}. \quad (1.9)$$

Such a prior distribution is known as a Jeffreys prior distribution.

Example 1.5

In constructing the prior in Example 1.1, did we assume (i) substantial prior knowledge; (ii) limited prior knowledge; or (iii) prior ignorance?



Limited prior knowledge – for mathematical convenience we used a conjugate prior, but we obtained the parameters in this prior through prior information collected from the epidemiologist.

Example 1.6

Suppose the parameter θ represents a probability. What might be the simplest prior to use here, if we are prior ignorant?



Rather than compute a Jeffreys prior, we might simply choose a Uniform $U(0, 1)$ prior here; that is, $\theta \sim U(0, 1)$, with

$$\pi(\theta) = 1.$$

This would give all outcomes across the permissible range of values for θ equal density.

Example 1.7

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known). Determine the posterior distribution assuming a vague prior for μ .



The conjugate prior distribution is a normal distribution. We have already seen that if the prior is $\mu \sim N(b, 1/d)$ then the posterior distribution is $\mu|\mathbf{x} \sim N(B, 1/D)$ where

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \quad \text{and} \quad D = d + n\tau.$$

If we now make our prior knowledge vague about μ by letting the prior variance tend to infinity ($d \rightarrow 0$), we obtain

$$B \rightarrow \bar{x} \quad \text{and} \quad D \rightarrow n\tau.$$

Therefore, assuming vague prior knowledge for μ results in a $N\{\bar{x}, 1/(n\tau)\}$ posterior distribution. Notice that the posterior mean is the sample mean (the likelihood mode) and that the posterior variance $1/(n\tau) \rightarrow 0$ as $n \rightarrow \infty$.

Example 1.8

Suppose we have a random sample from a Poisson distribution, that is, $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \dots, n$ (independent). Determine the posterior distribution assuming a vague prior for θ .



The conjugate prior distribution is a Gamma distribution. Recall that a $Ga(g, h)$ distribution has mean $m = g/h$ and variance $v = g/h^2$. Rearranging these formulae we obtain

$$g = \frac{m^2}{v} \quad \text{and} \quad h = \frac{m}{v}.$$

Clearly $g \rightarrow 0$ and $h \rightarrow 0$ as $v \rightarrow \infty$ (for fixed m). We have seen how taking a $Ga(g, h)$ prior distribution results in a $Ga(g + n\bar{x}, h + n)$ posterior distribution. Therefore, taking a vague prior distribution will give a $Ga(n\bar{x}, n)$ posterior distribution. Note that the posterior mean is $\bar{x}$ (the likelihood mode) and that the posterior variance $\bar{x}/n \rightarrow 0$ as $n \rightarrow \infty$.

Example 1.9

Suppose we have a random sample from an exponential distribution, that is, $X_i|\theta \sim \text{Exp}(\theta)$, $i = 1, 2, \dots, n$ (independent). Determine the Jeffreys prior for this model.



Recall that $L(\theta|\mathbf{x}) = f_{\mathbf{x}}(\mathbf{x}|\theta) = \theta^n e^{-n\bar{x}\theta}$, and therefore

$$\begin{aligned}\log L(\theta|\mathbf{x}) &= n \log \theta - n\bar{x}\theta \\ \Rightarrow \frac{\partial}{\partial \theta} \log L(\theta|\mathbf{x}) &= \frac{n}{\theta} - n\bar{x} \\ \Rightarrow \frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{x}) &= -\frac{n}{\theta^2} \\ \Rightarrow I(\theta) = E_{\mathbf{x}|\theta} \left[-\frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{X}) \right] &= \frac{n}{\theta^2}.\end{aligned}$$

Hence, the Jeffreys prior for this model is

$$\begin{aligned}\pi(\theta) &\propto I(\theta)^{1/2} \\ &\propto \frac{\sqrt{n}}{\theta}, \quad \theta > 0 \\ &\propto \frac{1}{\theta}, \quad \theta > 0.\end{aligned}$$

Notice that this distribution is improper since $\int_0^\infty d\theta/\theta$ is a divergent integral, and so we cannot find a constant which ensures that the density function integrates to one.

Notice also that this density is, in fact, a limiting form of a $Ga(g, h)$ density (ignoring the integration constant) since

$$\frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \propto \theta^{g-1} e^{-h\theta} \rightarrow \frac{1}{\theta}, \quad \text{as } g \rightarrow 0, \quad h \rightarrow 0.$$

Therefore, we obtain the same posterior distribution whether we adopt the Jeffreys prior or vague prior knowledge.

1.3 The asymptotic posterior distribution

If we have a statistical model $f(\mathbf{x}|\theta)$ for data $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, together with a prior distribution $\pi(\theta)$ for θ then

$$\sqrt{J(\hat{\theta})} (\theta - \hat{\theta})|\mathbf{x} \xrightarrow{\mathcal{D}} N(0, 1) \quad \text{as } n \rightarrow \infty,$$

where $\hat{\theta}$ is the likelihood mode and $J(\theta)$ is the observed information

$$J(\theta) = -\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{x}|\theta).$$

This means that, with increasing amounts of data, the posterior distribution looks more and more like a normal distribution. The result also gives us a useful approximation to the posterior distribution for θ when n is large:

$$\theta|\mathbf{x} \sim N\{\hat{\theta}, J(\hat{\theta})^{-1}\} \quad \text{approximately.}$$

Note that this limiting result is similar to one used in Frequentist statistics for the distribution of the maximum likelihood estimator, namely

$$\sqrt{I(\theta)} (\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, 1) \quad \text{as } n \rightarrow \infty,$$

where Fisher's information $I(\theta)$ is the expected value of the observed information, where the expectation is taken over the distribution of $\mathbf{X}|\theta$, that is, $I(\theta) = E_{\mathbf{X}|\theta}[J(\theta)]$. You may also have seen this result written as an approximation to the distribution of the maximum likelihood estimator in large samples, namely

$$\hat{\theta} \sim N\{\theta, I(\theta)^{-1}\} \quad \text{approximately.}$$

Example 1.10

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known). Determine the asymptotic posterior distribution for μ .

Recall that

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

and therefore

$$\begin{aligned} \log f(\mathbf{x}|\mu) &= \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2 \\ \Rightarrow \quad \frac{\partial}{\partial \mu} \log f(\mathbf{x}|\mu) &= -\frac{\tau}{2} \times \sum_{i=1}^n -2(x_i - \mu) = \tau \sum_{i=1}^n (x_i - \mu) = n\tau(\bar{x} - \mu) \\ \Rightarrow \quad \frac{\partial^2}{\partial \mu^2} \log f(\mathbf{x}|\mu) &= -n\tau \quad \Rightarrow \quad J(\mu) = -\frac{\partial^2}{\partial \mu^2} \log f(\mathbf{x}|\mu) = n\tau. \end{aligned}$$



... Solution to Example 1.10

We have

$$\begin{aligned} \frac{\partial}{\partial \mu} \log f(\mathbf{x}|\mu) = 0 & \implies \hat{\mu} = \bar{x} \\ & \implies J(\hat{\mu}) = n\tau \\ & \implies J(\hat{\mu})^{-1} = \frac{1}{n\tau}. \end{aligned}$$

Therefore, for large n , the (approximate) posterior distribution for μ is

$$\mu|\mathbf{x} \sim N\left(\bar{x}, \frac{1}{n\tau}\right).$$

Here the asymptotic posterior distribution is the same as the posterior distribution under vague prior knowledge.

1.4 Bayesian inference

The posterior distribution $\pi(\theta|\mathbf{x})$ summarises all our information about θ to date. However, sometimes it is helpful to reduce this distribution to a few key summary measures.

1.4.1 Estimation

Point estimates

There are many useful summaries for a typical value of a random variable with a particular distribution; for example, the mean, mode and median. The mode is used more often as a summary than is the case in frequentist statistics.

Confidence intervals/regions

A more useful summary of the posterior distribution is one which also reflects its variation. For example, a $100(1 - \alpha)\%$ *Bayesian confidence interval* for θ is any region C_α that satisfies $Pr(\theta \in C_\alpha|\mathbf{x}) = 1 - \alpha$. If θ is a continuous quantity with posterior probability density function $\pi(\theta|\mathbf{x})$ then

$$\int_{C_\alpha} \pi(\theta|\mathbf{x}) d\theta = 1 - \alpha.$$

The usual correction is made for discrete θ , that is, we take the largest region C_α such that $Pr(\theta \in C_\alpha | \mathbf{x}) \leq 1 - \alpha$. Bayesian confidence intervals are sometimes called *credible regions* or *plausible regions*. Clearly these intervals are not unique, since there will be many intervals with the correct probability coverage for a given posterior distribution.

A $100(1 - \alpha)\%$ *highest density interval* (HDI) for θ is the region

$$C_\alpha = \{\theta : \pi(\theta | \mathbf{x}) \geq \gamma\}$$

where γ is chosen so that $Pr(\theta \in C_\alpha | \mathbf{x}) = 1 - \alpha$. This region is sometimes called a *most plausible Bayesian confidence interval*. If the posterior distribution has many modes then it is possible that the HDI will be the union of several disjoint regions. Also, if the posterior distribution is unimodal (has one mode) and symmetric about its mean then the HDI is an equi-tailed interval, that is, takes the form $C_\alpha = (a, b)$, where $Pr(\theta < a | \mathbf{x}) = Pr(\theta > b | \mathbf{x}) = \alpha/2$; see Figure 1.3.

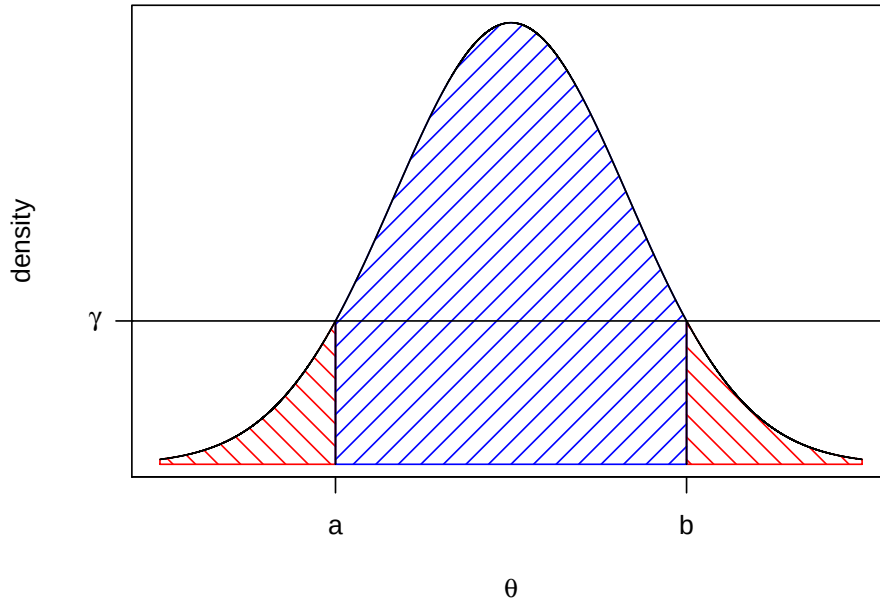


Figure 1.3: Construction of an HDI for a symmetric posterior density

Interpretation of confidence intervals/regions

Suppose C_B is a 95% Bayesian confidence interval for θ and C_F is a 95% frequentist confidence interval for θ . These intervals do not have the same interpretation:

- the probability that C_B contains θ is 0.95;
- the probability that C_F contains θ is either 0 or 1 — since θ does not have a (non-degenerate) probability distribution;
- the interval C_F covers the true value θ on 95% of occasions — in repeated applications of the formula.

Example 1.11

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ from a $N(\mu, 1/\tau)$ distribution (where τ is known). We have seen that, assuming vague prior knowledge, the posterior distribution is $\mu|\mathbf{x} \sim N\{\bar{x}, 1/(n\tau)\}$. Determine the $100(1 - \alpha)\%$ HDI for μ .



This distribution has a symmetric bell shape and so the HDI is an equi-tailed interval $C_\alpha = (a, b)$ with $Pr(\mu < a|\mathbf{x}) = \alpha/2$ and $Pr(\mu > b|\mathbf{x}) = \alpha/2$, that is,

$$a = \bar{x} - \frac{z_{\alpha/2}}{\sqrt{n\tau}} \quad \text{and} \quad b = \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n\tau}},$$

where z_α is the upper α -quantile of the $N(0, 1)$ distribution. For example, the 95% HDI for μ is

$$\left(\bar{x} - \frac{1.96}{\sqrt{n\tau}}, \bar{x} + \frac{1.96}{\sqrt{n\tau}} \right).$$

Note that this interval is numerically identical to the 95% frequentist confidence interval for the (population) mean of a normal random sample with known variance. However, the interpretation is very different.

Example 1.12

Recall Example 1.1 on the number of cases of foodborne botulism in England and Wales. The data were modelled as a random sample from a Poisson distribution with mean θ . Using a $Ga(2, 1)$ prior distribution, we found the posterior distribution to be $\theta|\mathbf{x} \sim Ga(8, 9)$ (see Figure 1.5). Determine the $100(1 - \alpha)\%$ HDI for θ .

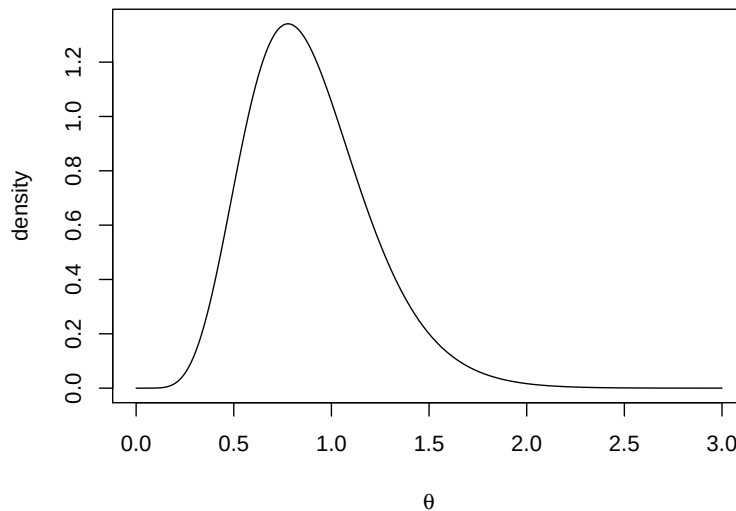


Figure 1.4: Posterior density for θ



... *Solution to Example 1.12*

The HDI must take the form $C_\alpha = (a, b)$ if it is to include the values of θ with the highest probability density. Suppose that $F(\cdot)$ and $f(\cdot)$ are the posterior distribution and density functions. Then the end-points a and b must satisfy

$$Pr(a < \theta < b | \mathbf{x}) = F(b) - F(a) = 1 - \alpha$$

and

$$\pi(\theta = a | \mathbf{x}) - \pi(\theta = b | \mathbf{x}) = f(a) - f(b) = 0.$$

Unfortunately, there is no simple analytical solution to these equations and so numerical methods have to be employed to determine a and b . However, if we have the quantile function $F^{-1}(\cdot)$ for $\theta | \mathbf{x}$ then we can find the solution by noticing that we can write b in terms of a :

$$b = F^{-1}\{1 - \alpha + F(a)\},$$

for $0 < a < F^{-1}(\alpha)$. Therefore, we can determine the correct choice of a by minimising the function

$$g(a) = (f(a) - f[F^{-1}\{1 - \alpha + F(a)\}])^2.$$

within the range $0 < a < F^{-1}(\alpha)$. The values of a and b can be determined using the optimizer function `optimize` in R.

The R package `nclbayes` contains functions to determine the HDI for several distributions. The function for the Gamma distribution is `hdiGamma` and we can calculate the 95% HDI for the $Ga(8, 9)$ posterior distribution by using the commands

```
install.packages("nclbayes", repos="http://R-Forge.R-project.org")
library(nclbayes)
hdiGamma(p=0.95, a=8, b=9)
```

Taking $1 - \alpha = 0.95$ and using such R code gives $a = 0.3304362$ and $b = 1.5146208$. To check this answer, R gives $Pr(a < \theta < b | \mathbf{x}) = 0.95$, $\pi(\theta = b | \mathbf{x}) = 0.1877215$ and $\pi(\theta = a | \mathbf{x}) = 0.1877427$. Thus the 95% HDI is $(0.3304362, 1.514621)$.

The package also has functions `hdiBeta` for the Beta distribution and `hdiInvchi` for the Inv-Chi distribution (introduced in Chapter 2).

1.4.2 Prediction

Much of statistical inference (both Frequentist and Bayesian) is aimed towards making statements about a parameter θ . Often the inferences are used as a yardstick for similar future experiments. For example, we may want to predict the outcome when the experiment is performed again.

Clearly there will be uncertainty about the future outcome of an experiment. Suppose this future outcome Y is described by a probability (density) function $f(y|\theta)$. There are several ways we could make inferences about what values of Y are likely. For example, if we have an estimate $\hat{\theta}$ of θ we might base our inferences on $f(y|\theta = \hat{\theta})$. Obviously this is not the best we can do; such inferences ignore the fact that it is unlikely that $\theta = \hat{\theta}$.

Implicit in the Bayesian framework is the concept of the *predictive distribution*. This distribution describes how likely are different outcomes of a future experiment. The predictive probability (density) function is calculated as

$$f(y|\mathbf{x}) = \int_{\Theta} f(y|\theta) \pi(\theta|\mathbf{x}) d\theta$$

when θ is a continuous quantity. From this equation, we can see that the predictive distribution is formed by weighting the possible values of θ in the future experiment $f(y|\theta)$ by how likely we believe they are to occur $\pi(\theta|\mathbf{x})$.

If the true value of θ were known, say θ_0 , then any prediction can do no better than one based on $f(y|\theta = \theta_0)$. However, as (generally) θ is unknown, the predictive distribution is used as the next best alternative.

We can use the predictive distribution to provide a useful range of plausible values for the outcome of a future experiment. This *prediction interval* is similar to a HDI interval. A $100(1 - \alpha)\%$ *prediction interval* for Y is the region $C_{\alpha} = \{y : f(y|\mathbf{x}) \geq \gamma\}$ where γ is chosen so that $Pr(Y \in C_{\alpha}|\mathbf{x}) = 1 - \alpha$.

Example 1.13

Recall Example 1.1 on the number of cases of foodborne botulism in England and Wales. The data for 1998–2005 were modelled by a Poisson distribution with mean θ . Using a $Ga(2, 1)$ prior distribution, we found the posterior distribution to be $\theta|\mathbf{x} \sim Ga(8, 9)$. Determine the predictive distribution for the number of cases for the following year (2006).



Suppose the number of cases in 2006 is Y , with $Y|\theta \sim Po(\theta)$. The predictive probability function of Y is, for $y = 0, 1, \dots$

$$f(y|\mathbf{x}) = \int_{\Theta} f(y|\theta) \pi(\theta|\mathbf{x}) d\theta = \int_0^{\infty} \frac{\theta^y e^{-\theta}}{y!} \times \frac{9^8 \theta^7 e^{-9\theta}}{\Gamma(8)} d\theta$$

 ... Solution to Example 1.13

$$\begin{aligned}
 &= \frac{9^8}{y! \Gamma(8)} \int_0^\infty \theta^{y+7} e^{-10\theta} d\theta \\
 &= \frac{9^8}{y! \Gamma(8)} \times \frac{\Gamma(y+8)}{10^{y+8}} \\
 &= \frac{(y+7)!}{y! 7!} \times 0.9^8 \times 0.1^y \\
 &= \binom{y+7}{7} \times 0.9^8 \times 0.1^y.
 \end{aligned}$$

You may not recognise this probability function but it is related to that of a negative binomial distribution. Suppose $Z \sim \text{NegBin}(r, p)$ with probability function

$$Pr(Z = z) = \binom{z-1}{r-1} p^r (1-p)^{z-r}, \quad z = r, r+1, \dots$$

Then $W = Z - r$ has probability function

$$Pr(W = w) = Pr(Z = w + r) = \binom{w+r-1}{r-1} p^r (1-p)^w, \quad w = 0, 1, \dots$$

This is the same probability function as our predictive probability function, with $r = 8$ and $p = 0.9$. Therefore $Y|\mathbf{x} \sim \text{NegBin}(8, 0.9) - 8$. Note that, unfortunately R also calls the distribution of W a negative binomial distribution with parameters r and p . To distinguish between this distribution and the $\text{NegBin}(r, p)$ distribution used above, we shall denote the distribution of W as a $\text{NegBin}_R(r, p)$ distribution – it has mean $r(1-p)/p$ and variance $r(1-p)/p^2$. Thus $Y|\mathbf{x} \sim \text{NegBin}_R(8, 0.9)$.

We can compare this predictive distribution with a naive predictive distribution based on an estimate of θ . Here we shall base our naive predictive distribution on the maximum likelihood estimate $\hat{\theta} = 0.75$, that is, use the distribution $Y|\theta = \hat{\theta} \sim \text{Po}(0.75)$. Thus, the naive predictive probability function is

$$f(y|\theta = \hat{\theta}) = \frac{0.75^y e^{-0.75}}{y!}, \quad y = 0, 1, \dots$$

Numerical values for the predictive and naive predictive probability functions are given in Table 1.5.

Again, the naive predictive distribution is a predictive distribution which, instead of using the correct posterior distribution, uses a degenerate posterior distribution $\pi^*(\theta|\mathbf{x})$ which essentially allows only one value: $Pr_{\pi^*}(\theta = 0.75|\mathbf{x}) = 1$ and standard deviation $SD_{\pi^*}(\theta|\mathbf{x}) = 0$. Note that the correct posterior standard deviation of θ is $SD_{\pi}(\theta|\mathbf{x}) =$

	correct	naive
y	$f(y \mathbf{x})$	$f(y \theta = \hat{\theta})$
0	0.430	0.472
1	0.344	0.354
2	0.155	0.133
3	0.052	0.033
4	0.014	0.006
5	0.003	0.001
≥ 6	0.005	0.002

Table 1.5: Predictive and naive predictive probability functions

$\sqrt{8}/9 = 0.314$. Using a degenerate posterior distribution results in the naive predictive distribution having too small a standard deviation:

$$SD(Y|X = 1) = \begin{cases} 0.994 & \text{using the correct } \pi(\theta|\mathbf{x}) \\ 0.866 & \text{using the naive } \pi^*(\theta|\mathbf{x}), \end{cases}$$

these values being calculated from $NegBin_R(8, 0.9)$ and $Po(0.75)$ distributions.

Using the numerical table of predictive probabilities, we can see that $\{0, 1, 2\}$ is a 92.9% prediction set/interval. This is to be contrasted with the more “optimistic” calculation using the naive predictive distribution which shows that $\{0, 1, 2\}$ is a 95.9% prediction set/interval.

Candidate's formula

In the previous example, a non-trivial integral had to be evaluated. However, when the past data $\mathbf{x}$ and future data y are independent (given θ) and we use a conjugate prior distribution, another (easier) method can be used to determine the predictive distribution.

Using Bayes Theorem, the posterior density for θ given $\mathbf{x}$ and y is

$$\begin{aligned} \pi(\theta|\mathbf{x}, y) &= \frac{\pi(\theta)f(\mathbf{x}, y|\theta)}{f(\mathbf{x}, y)} \\ &= \frac{\pi(\theta)f(\mathbf{x}|\theta)f(y|\theta)}{f(\mathbf{x})f(y|\mathbf{x})} \quad \text{since } \mathbf{X} \text{ and } Y \text{ are independent given } \theta \\ &= \frac{\pi(\theta|\mathbf{x})f(y|\theta)}{f(y|\mathbf{x})}. \end{aligned}$$

Rearranging, we obtain

$$f(y|\mathbf{x}) = \frac{f(y|\theta)\pi(\theta|\mathbf{x})}{\pi(\theta|\mathbf{x}, y)}.$$

This is known as Candidate's formula. The right-hand-side of this equation looks as if it depends on θ but, in fact, any terms in θ will be cancelled between the numerator and denominator.

Example 1.14

Rework Example 1.13 using Candidate's formula to determine the number of cases in 2006.



Let Y denote the number of cases in 2006. We know that $\theta|\mathbf{x} \sim Ga(8, 9)$ and $Y|\theta \sim Po(\theta)$. Using Example 1.2 we obtain

$$\theta|\mathbf{x}, y \sim Ga(8 + y, 10).$$

Therefore the predictive probability function of Y is, for $y = 0, 1, \dots$

$$\begin{aligned} f(y|\mathbf{x}) &= \frac{f(y|\theta) \pi(\theta|\mathbf{x})}{\pi(\theta|\mathbf{x}, y)} \\ &= \frac{\frac{\theta^y e^{-\theta}}{y!} \times \frac{9^8 \theta^7 e^{-9\theta}}{\Gamma(8)}}{\frac{10^{8+y} \theta^{7+y} e^{-10\theta}}{\Gamma(8+y)}} \\ &= \frac{\Gamma(8+y)}{y! \Gamma(8)} \times \frac{9^8}{10^{8+y}} \\ &= \frac{(y+7)!}{y! 7!} \times 0.9^8 \times 0.1^y \\ &= \binom{y+7}{7} \times 0.9^8 \times 0.1^y. \end{aligned}$$

1.5 Sequential Bayesian updating

Throughout MAS2903, and this review chapter in the current module so far, we have focussed on using Bayes Theorem to perform an update of our beliefs about the parameter θ after having observed a single sample of data $\mathbf{x}$.

In many real-world problems in data science, new samples of data are being observed continuously; hence, it might be reasonable to assume our beliefs about θ could also be continuously updated. In such a setting, once the first sample of data $\mathbf{x}^{(1)}$ has been used to update our beliefs about θ , the resulting posterior distribution $\pi(\theta|\mathbf{x}^{(1)})$ represents our current beliefs about θ ; if we were to observe a new sample $\mathbf{x}^{(2)}$ we might then use $\pi(\theta|\mathbf{x}^{(1)})$ as our “new prior”, and update once again to obtain the “new posterior” $\pi(\theta|\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$.

In other words, we might replace the usual mantra

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

with

$$\text{Revised} \propto \text{Current} \times \text{New likelihood},$$

with the ability to update our likelihood function with each new sample of data. Why does this work?

Suppose we have data assumed drawn from a probability distribution with parameter θ .

- Before we make any observations at all, our prior beliefs about θ are represented by $\pi(\theta)$. We might refer to this as our “original prior”.
- Now we observe a sample $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$; our posterior is then

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta) \times f(\mathbf{x}|\theta) = \pi(\theta) \times \prod_{i=1}^n f(x_i|\theta),$$

assuming our observations are a random sample.

- Now suppose we make a new observation x_{n+1} assumed from the same probability distribution as $\mathbf{x}$.
 - Before making this observation, our *current* beliefs about θ are represented by $\pi(\theta|\mathbf{x})$; re-label this $\pi_n(\theta)$, our “new prior” after having made the initial n observations.
 - Assuming the new observation is drawn from the same probability model as the sample $\mathbf{x}$, then the *new likelihood* has the same functional form as the likelihood for $\mathbf{x}$

– Therefore,

$$\begin{aligned}
 \text{Revised} &\propto \text{Current} \times \text{New likelihood} \\
 \pi_{n+1}(\theta|x_{n+1}) &\propto \pi_n(\theta) \times f(x_{n+1}|\theta) \\
 &\propto \pi(\theta|\mathbf{x}) \times f(x_{n+1}|\theta) \\
 &\propto \left\{ \pi(\theta) \times \prod_{i=1}^n f(x_i|\theta) \right\} \times f(x_{n+1}|\theta) \\
 &\propto \pi(\theta) \times \prod_{i=1}^{n+1} f(x_i|\theta) \\
 &\propto \text{original prior} \times \text{full likelihood},
 \end{aligned}$$

where the “full likelihood” just represents the likelihood of all our observations together, i.e. the vector $(x_1, x_2, \dots, x_n, x_{n+1})^T$.

Example 1.15

Table 1.6 shows provides an update to table 1.1, giving the number of cases of foodbourne botulism in England and Wales between 2006 and 2010.

Use your solution to Example 1.1, and sequential Bayesian updating, to obtain your new posterior distribution for θ .

Year	2006	2007	2008	2009	2010
Cases	1	0	0	1	1

Table 1.6: Number of cases of foodbourne botulism in England and Wales, 2006–2010



Our current beliefs are represented by the “old posterior” – from Example 1.1, we found that $\theta|\mathbf{x} \sim \text{Ga}(8, 9)$, and so

$$\text{Current} = \frac{9^8 \theta^7 e^{-9\theta}}{7!} \propto \theta^7 e^{-9\theta}.$$

We also have

$$\begin{aligned}
 \text{New likelihood} &= e^{-\theta} \theta^1 \times e^{-\theta} \theta^0 \times \dots \times e^{-\theta} \theta^1 \\
 &= e^{-5\theta} \theta^{1+0+\dots+1} \propto e^{-5\theta} \theta^3.
 \end{aligned}$$

Therefore

$$\begin{aligned}
 \text{Revised} &\propto \theta^7 e^{-9\theta} \times \theta^3 e^{-5\theta} \\
 &= \theta^{10} e^{-14\theta},
 \end{aligned}$$

and so it must be the case that our revised posterior is $\theta|\mathbf{x} \sim \text{Ga}(11, 14)$. This gives the further updates as shown in Table 1.7.

From the information in Table 1.7, we see that – in light of this new sample of data – we are more sure about θ , with a further reduction in $SD(\theta)$. We can see this in the updated density plots in Figure 1.5, with a more peaked and concentrated posterior. Before having collected any data, and based on the epidemiologist's prior information only, in a ten year period we might have expected around 20 cases of foodbourne botulism (taking the expected value of around 2 per year); after the first sample of data, this reduces to around 18 cases, and now to just under 16 cases in light of the updated dataset.

	Prior: $Ga(2, 1)$ (Original prior)	Posterior: $Ga(8, 9)$ (New prior)	Revised: $Ga(11, 14)$ (New posterior)
$Mode(\theta)$	1.00	0.78	0.71
$E(\theta)$	2.00	0.89	0.79
$SD(\theta)$	1.41	0.31	0.24

Table 1.7: Changes in beliefs about θ after second update

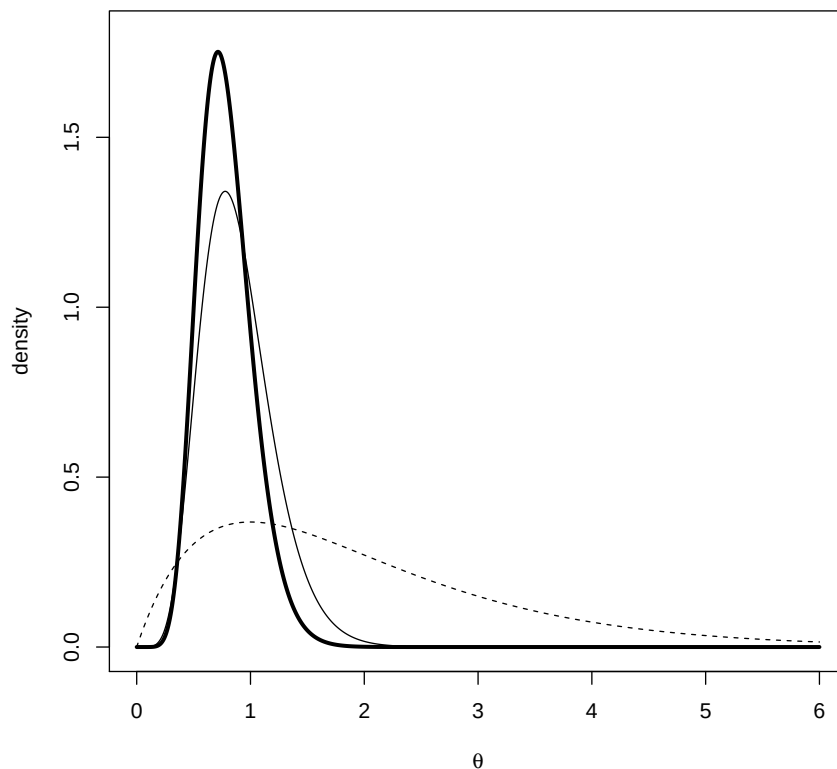


Figure 1.5: Prior (dashed) and posterior (solid) densities for θ , with the heavy line representing the posterior density after the second update

Example 1.16

Devise a generic sequential Bayesian updating plan for successive random samples from a Poisson distribution, assuming a conjugate prior specification.



Suppose our first sample is $\mathbf{x}^{(1)}$ with sample size n_1 , that is, $\mathbf{x}^{(1)} = (x_1, x_2, \dots, x_{n_1})^T$. From earlier examples, we know that the likelihood function is

$$f(\theta|\mathbf{x}^{(1)}) \propto e^{-n_1\theta} \theta^{n_1\bar{x}^{(1)}}.$$

Therefore

$$\pi(\theta|\mathbf{x}^{(1)}) \propto \theta^{g-1} e^{-h\theta} \times e^{-n_1\theta} \theta^{n_1\bar{x}^{(1)}} = \theta^{(g+n_1\bar{x}^{(1)})-1} e^{-(h+n_1)\theta},$$

and so $\theta|\mathbf{x}^{(1)} \sim Ga(g + n_1\bar{x}^{(1)}, h + n_1)$. Now suppose we have another sample $\mathbf{x}^{(2)}$ with sample size n_2 ; then

$$\begin{aligned} \pi(\theta|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) &\propto \theta^{(g+n_1\bar{x}^{(1)})-1} e^{-(h+n_1)\theta} \times e^{-n_2\theta} \theta^{n_2\bar{x}^{(2)}} \\ &\propto \theta^{(g+n_1\bar{x}^{(1)}+n_2\bar{x}^{(2)})-1} e^{-(h+n_1+n_2)\theta}, \end{aligned}$$

giving $\theta|\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \sim Ga(g + n_1\bar{x}^{(1)} + n_2\bar{x}^{(2)}, h + n_1 + n_2)$. Continuing this way, we can see that a generic sequential updating scheme would give rise to the posterior:

$$\theta|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots \sim Ga(g + n_1\bar{x}^{(1)} + n_2\bar{x}^{(2)} + \dots, h + n_1 + n_2 + \dots);$$

that is, our posterior will always be $Ga(G, H)$, and at each update j G will increase by $n_j\bar{x}^{(j)}$ and H will increase by n_j . The parameters in the posterior can be expressed recursively:

$$\begin{aligned} G_j &= G_{j-1} + n_j\bar{x}^{(j)} \quad \text{and} \\ H_j &= H_{j-1} + n_j. \end{aligned}$$

1.6 Learning objectives

By the end of this chapter, you should be able to:

- Determine the likelihood function using a random sample from **any** distribution
- Combine this likelihood function with **any** prior distribution to obtain the posterior distribution
- Name the posterior distribution if it is a “standard” distribution listed in these notes or on the exam paper – this list may well include distributions that are standard within the subject but which you have not met before. If the posterior distribution is not a “standard” distribution then it is okay just to give its density (or probability function) up to a constant.
- Do all the above for a particular data set or for a general case with random sample $x_1, \dots, x_n$
- Describe the different levels of prior information; determine and use conjugate priors and vague priors
- Determine the asymptotic posterior distribution
- Determine the predictive distribution, particularly when having a random sample from any distribution and a conjugate prior via Candidate’s formula
- Describe and calculate the confidence intervals, HDIs and prediction intervals
- Operate sequential Bayesian updating schemes

