MAS3902

Bayesian Inference

Semester 1, 2019–20

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. It also introduces the idea of a *mixture prior distribution*.
- **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 module MAS8951: Modern Bayesian Inference.

Contents

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

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

where

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

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

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\propto \pi(\theta) \times f(\boldsymbol{x}|\theta) \\ i.e. \quad \text{posterior} &\propto \text{prior} \times \text{likelihood}. \end{aligned}$$

Example 1.1

Table 1.1 shows some data on the number of cases of foodbourne 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 foodbourne botulism in England and Wales, 1998–2005

An expert in the epidemiology of similar diseases gives their prior distribution for the rate θ as a $Ga(2,1)$ distribution, with density

$$\pi(\theta) = \theta\, e^{-\theta}, \quad \theta > 0, \tag{1.1}$$

and mean $E(\theta) = 2$ and variance $Var(\theta) = 2$. Determine the posterior distribution for θ.

Solution

The data are observations on $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, 8$ (independent). Therefore, the likelihood function for θ is

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{8} \frac{\theta^{x_i} e^{-\theta}}{x_i!}, \qquad \theta > 0 \\ &= \frac{\theta^{2+0+\cdots+1} e^{-8\theta}}{2! \times 0! \times \cdots \times 1!}, \qquad \theta > 0 \\ &= \frac{\theta^6 e^{-8\theta}}{4}, \qquad \theta > 0. \end{aligned} \tag{1.2}$$

Bayes Theorem combines the expert opinion with the observed data, and gives the posterior density function as

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\ &\propto \theta e^{-\theta} \times \frac{\theta^6 e^{-8\theta}}{4}, \qquad \theta > 0 \\ &= k\, \theta^7 e^{-9\theta}, \qquad \theta > 0. \end{aligned} \tag{1.3}$$

The only continuous distribution with density of the form $k\theta^{g-1}e^{-h\theta}$, $\theta > 0$ is the $Ga(g,h)$ distribution. Therefore, the posterior distribution must be $\theta|\boldsymbol{x} \sim Ga(8,9)$.

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

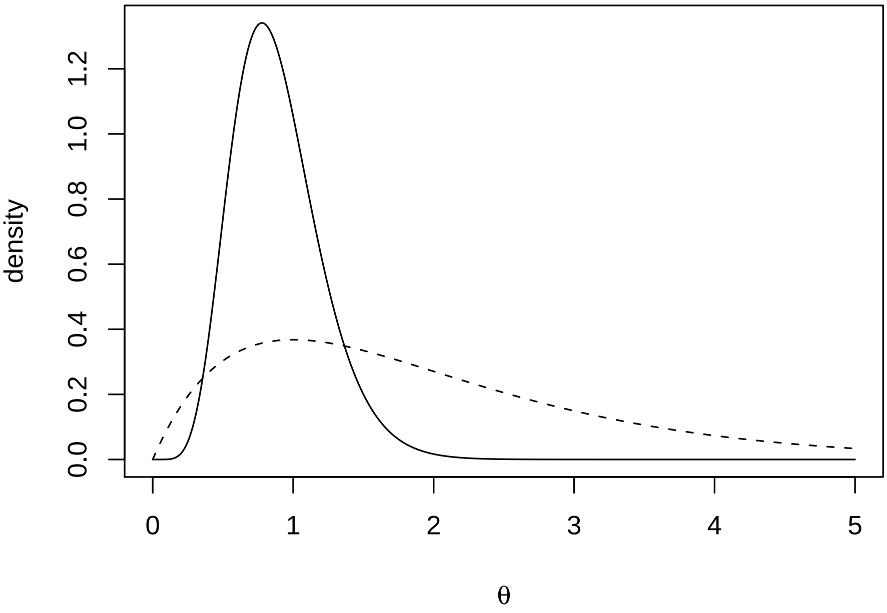

Figure 1.1: Prior (dashed) and posterior (solid) densities for θ

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.

	Prior	Likelihood	Posterior
	(1.1)	(1.2)	(1.3)
$Mode(\theta)$	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 $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent) and our prior beliefs about θ are summarised by a $Ga(g, h)$ distribution (with g and h known), with density

$$\pi(\theta) = \frac{h^g\,\theta^{g-1}e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \tag{1.4}$$

Determine the posterior distribution for θ.

Solution

The likelihood function for θ is

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}, \qquad \theta > 0 \\ &\propto \theta^{n\bar{x}} e^{-n\theta}, \qquad \theta > 0. \end{aligned} \tag{1.5}$$

Using Bayes Theorem, the posterior density function is

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\ &\propto \frac{h^g\, \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \theta^{n\bar{x}} e^{-n\theta}, \qquad \theta > 0 \\ \text{i.e.} \quad \pi(\theta|\boldsymbol{x}) &= k\theta^{g+n\bar{x}-1} e^{-(h+n)\theta}, \qquad \theta > 0 \end{aligned} \tag{1.6}$$

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 $\theta|\boldsymbol{x} \sim Ga(G = g + n\bar{x}, H = h + n)$.

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 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,

	Prior	Likelihood	Posterior
	(1.4)	(1.5)	(1.6)
$Mode(\theta)$	$(g-1)/h$	$\bar{x}$	$(g+n\bar{x}-1)/(h+n)$
$E(\theta)$	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

$$E(\theta|\boldsymbol{x}) > E(\theta) \quad\Longleftrightarrow\quad Mode_\theta\{f(\boldsymbol{x}|\theta)\} > E(\theta).$$

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

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

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/variance) of the distribution rather than the variance itself. So we will assume that this population has unknown mean μ but known precision τ: $X_i|\mu \sim N(\mu,1/\tau)$, $i=1,2,\ldots,n$ (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

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu-b)^2\right\}. \tag{1.7}$$

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.

Solution

The likelihood function for μ is

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

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

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

Therefore

$$f(\boldsymbol{x}|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}. \tag{1.8}$$

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

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

as c does not depend on μ. Let

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \qquad \text{and} \qquad D = d + n\tau. \tag{1.9}$$

Then

$$\pi(\mu|\boldsymbol{x}) = k \exp\left\{-\frac{D}{2}(\mu - B)^2\right\}, \tag{1.10}$$

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 $\mu|\boldsymbol{x} \sim N(B, 1/D)$.

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

	Prior	Likelihood	Posterior
	(1.7)	(1.8)	(1.10)
$Mode(\mu)$	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

$$E(\mu|\boldsymbol{x}) > E(\mu) \qquad \Longleftrightarrow \qquad Mode_{\mu}\{f(\boldsymbol{x}|\mu)\} > E(\mu).$$

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.

Solution

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 $n = 23$. Therefore, using Example 1.3, the posterior distribution is $\mu|\boldsymbol{x} \sim N(B, 1/D)$, where

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

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 $\mu|\boldsymbol{x} \sim N(5.484, 0.0415^2)$ and so

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

$$Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9510,$$

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

calculated using the R command `pnorm(5.615,5.484,0.0415)-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

$$Pr(5.315 < \mu < 5.715) = 0.1896,$$

calculated using the R command `pnorm(5.615,5.4,0.4)-pnorm(5.415,5.4,0.4)`.

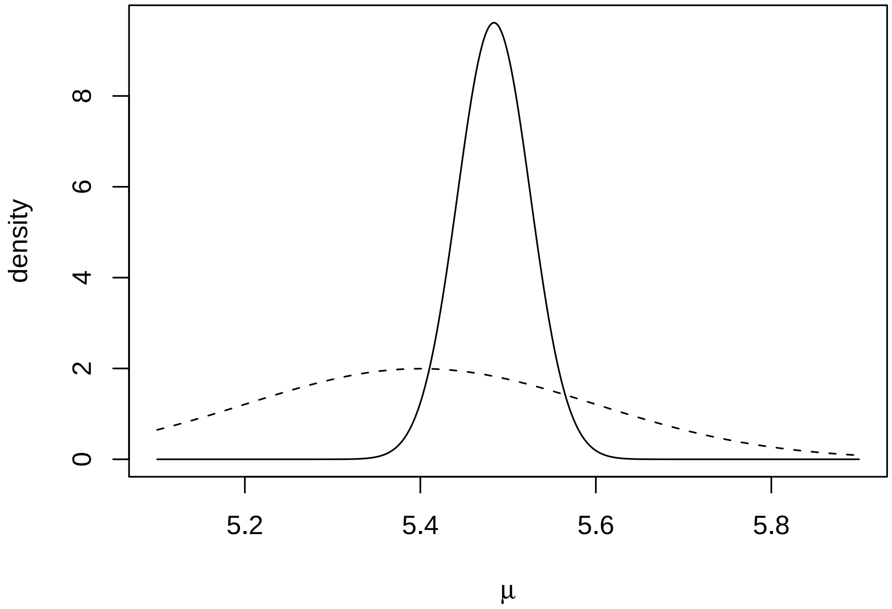

Figure 1.2: Prior (dashed) and posterior (solid) densities for the earth's density

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 $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$.

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

Notice that the conjugate family depends crucially on the model chosen for the data $\boldsymbol{x}$. 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

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 $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$.

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

Example 1.5

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.

Solution

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 $\mu|\boldsymbol{x} \sim N(B, 1/D)$ where

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

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

$$B \to \bar{x} \qquad \text{and} \qquad D \to n\tau.$$

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 $1/(n\tau) \to 0$ as $n \to \infty$.

Example 1.6

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

Solution

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

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

Clearly $g \to 0$ and $h \to 0$ as $v \to \infty$ (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 $\bar{x}/n \to 0$ and $n \to \infty$.

1.3 Asymptotic posterior distribution

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

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

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

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

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:

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

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

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

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

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

Example 1.7

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

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

and therefore

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

Solution

We have

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

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

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

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

1.4 Bayesian inference

The posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ summarises all our information about $\boldsymbol{\theta}$ 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 $Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$. If θ is a continuous quantity with posterior probability density function $\pi(\theta|\boldsymbol{x})$ then

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

The usual correction is made for discrete θ, that is, we take the largest region C_α such that $Pr(\theta \in C_\alpha|\boldsymbol{x}) \leq 1-\alpha$. 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

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

where γ is chosen so that $Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$. 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 $Pr(\theta < a|\boldsymbol{x}) = Pr(\theta > b|\boldsymbol{x}) = \alpha/2$; see Figure 1.3.

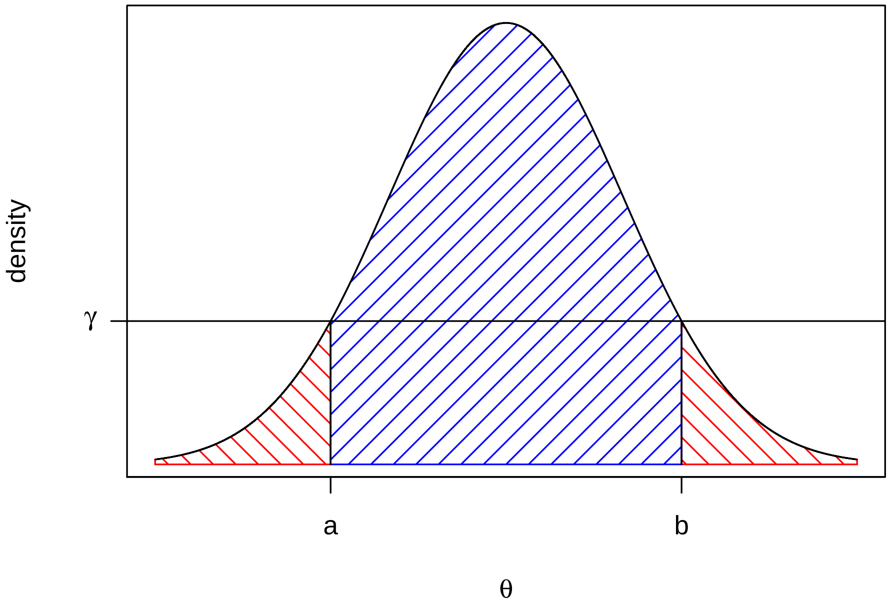

Figure 1.3: Construction of an HDI for a symmetric posterior density

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 1.8

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

Solution

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

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

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 1.9

Recall Example 1.1 on the number of cases of foodbourne 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 $\theta|\boldsymbol{x} \sim Ga(8,9)$. This posterior density is shown in Figure 1.4. Determine the $100(1-\alpha)\%$ HDI for θ.

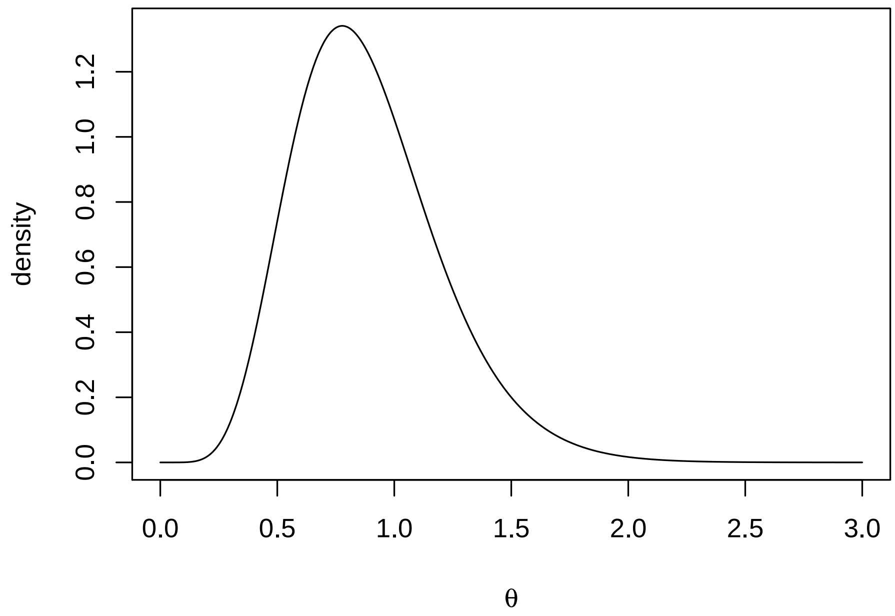

Figure 1.4: Posterior density for θ

Solution

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

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

and

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

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 $\theta|\boldsymbol{x}$ 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

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

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

```
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 $Pr(a < \theta < b|\boldsymbol{x}) = 0.95$, $\pi(\theta = b|\boldsymbol{x}) = 0.1877215$ and $\pi(\theta = a|\boldsymbol{x}) = 0.1877427$. 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, as such inferences ignore the fact that it is very 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

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

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 $\pi(\theta|\boldsymbol{x})$.

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 $C_\alpha = \{y : f(y|\boldsymbol{x}) \geq \gamma\}$ where γ is chosen so that $Pr(Y \in C_\alpha|\boldsymbol{x}) = 1 - \alpha$.

Example 1.10

Recall Example 1.1 on the number of cases of foodbourne 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 $\theta|\boldsymbol{x} \sim Ga(8,9)$. Determine the predictive distribution for the number of cases for the following year (2006).

Solution

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

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

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

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

Then $W = Z - r$ has probability function

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

This is the same probability function as our predictive probability function, with $r = 8$ and $p = 0.9$. Therefore $Y|\boldsymbol{x} \sim NegBin(8, 0.9) - 8$. 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 $NegBin(r,p)$ distribution used above, we shall denote the distribution of W as a $NegBin_{\mathtt{R}}(r,p)$ distribution – it has mean $r(1-p)/p$ and variance $r(1-p)/p^2$. Thus $Y|\boldsymbol{x} \sim NegBin_{\mathtt{R}}(8, 0.9)$.

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 $Y|\theta = \hat{\theta} \sim Po(0.75)$. Thus, the naive predictive probability function is

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

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

	correct	naive
y	$f(y\|\boldsymbol{x})$	$f(y\|\theta = \hat{\theta})$
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

Again, the naive predictive distribution is a predictive distribution which, instead of using the correct posterior distribution, uses a degenerate posterior distribution $\pi^*(\theta|\boldsymbol{x})$ which essentially allows only one value: $Pr_{\pi^*}(\theta = 0.75|\boldsymbol{x}) = 1$ and standard deviation $SD_{\pi^*}(\theta|\boldsymbol{x}) = 0$. Note that the correct posterior standard deviation of θ is $SD_{\pi}(\theta|\boldsymbol{x}) = \sqrt{8}/9 = 0.314$. Using a degenerate posterior distribution results in the naive predictive distribution having too small a standard deviation:

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

these values being calculated from $NegBin_{\mathrm{R}}(8, 0.9)$ 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 $\boldsymbol{x}$ 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 $\boldsymbol{x}$ and y is

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

Rearranging, we obtain

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

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 1.11

Rework Example 1.10 using Candidate's formula to determine the number of cases in 2006.

Solution

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

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

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

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

1.5 Mixture prior distributions

Sometimes prior beliefs cannot be adequately represented by a simple distribution, for example, a normal distribution or a beta distribution. In such cases, mixtures of distributions can be useful.

Example 1.12

Investigations into infants suffering from severe *idiopathic respiratory distress syndrome* have shown that whether the infant survives may be related to their weight at birth. Suppose that you are interested in developing a prior distribution for the mean birth weight μ of such infants. You might have a normal $N(2.3, 0.52^2)$ prior distribution for the mean birth weight (in kg) of infants who survive and a normal $N(1.7, 0.66^2)$ prior distribution for infants who die. If you believe that the proportion of infants that survive is 0.6, what is your prior distribution of birth weights of infants suffering from this syndrome?

Solution

Let $T = 1, 2$ denote whether the infant survives or dies. Then the information above tells us

$$\mu|T = 1 \sim N(2.3, 0.52^2) \quad \text{and} \quad \mu|T = 2 \sim N(1.7, 0.66^2).$$

In terms of density functions, we have

$$\pi(\mu|T = 1) = \phi\left(\mu|2.3, 0.52^2\right) \quad \text{and} \quad \pi(\mu|T = 2) = \phi\left(\mu|1.7, 0.66^2\right),$$

where $\phi(\cdot|a, b^2)$ is the normal $N(a, b^2)$ density function.

The prior distribution of birth weights of infants suffering from this syndrome is the (marginal) distribution of μ. Using the Law of Total Probability, the marginal density of μ is

$$\begin{aligned}\pi(\mu) &= Pr(T = 1) \times \pi(\mu|T = 1) + Pr(T = 2) \times \pi(\mu|T = 2)\\ &= 0.6\,\phi\left(\mu|2.3, 0.52^2\right) + 0.4\,\phi\left(\mu|1.7, 0.66^2\right).\end{aligned}$$

We write this as

$$\mu \sim 0.6\,N(2.3, 0.52^2) + 0.4\,N(1.7, 0.66^2).$$

This prior distribution is a mixture of two normal distributions. Figure 1.5 shows the overall (mixture) prior distribution $\pi(\mu)$ and the "component" distributions describing prior beliefs about the mean weights of those who survive and those who die. Notice that, in this example, although the mixture distribution is a combination of two distributions, each with one mode, this mixture distribution has only one mode. Also, although the component distributions are symmetric, the mixture distribution is not symmetric.

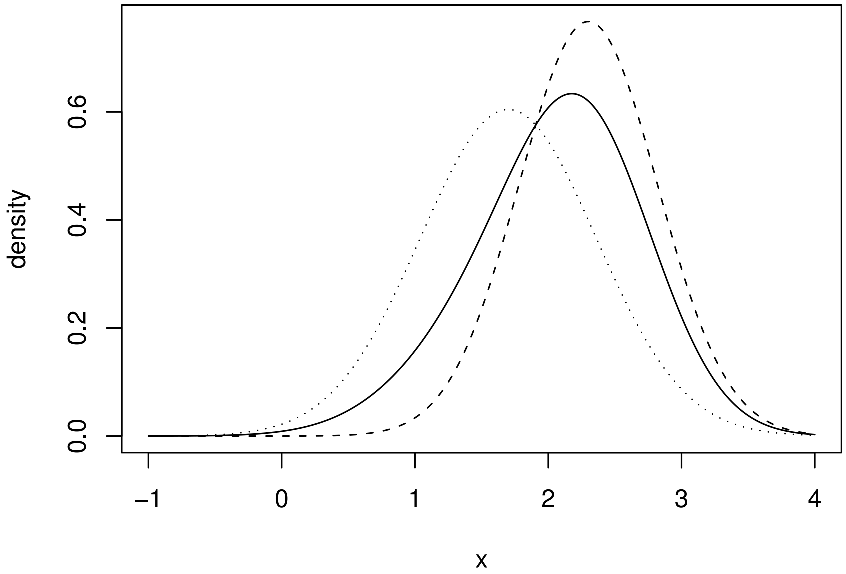

Figure 1.5: Plot of the mixture density (solid) with its component densities (survive – dashed; die – dotted)

Definition 1.2

A *mixture* of the distributions $\pi_i(\theta)$ with *weights* p_i $(i = 1, 2, \ldots, m)$ has probability (density) function

$$\pi(\theta) = \sum_{i=1}^{m} p_i \pi_i(\theta). \tag{1.11}$$

Figure 1.6 contains a plot of two quite different mixture distributions. One mixture distribution has a single mode and the other has two modes. In general, a mixture distribution whose m component distributions each have a single mode will have at most m modes.

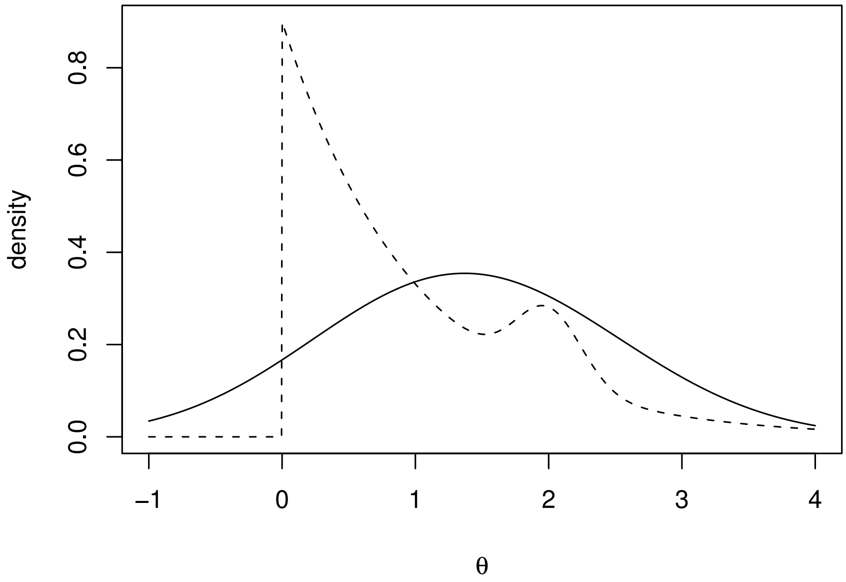

Figure 1.6: Plot of two mixture densities: solid is $0.6N(1,1)+0.4N(2,1)$; dashed is $0.9Exp(1)+0.1N(2,0.25^2)$

In order for a mixture distribution to be proper, we must have

$$\begin{aligned}1 &= \int_\Theta \pi(\theta)\,d\theta\\ &= \int_\Theta \sum_{i=1}^m p_i\pi_i(\theta)\,d\theta\\ &= \sum_{i=1}^m p_i \int_\Theta \pi_i(\theta)\,d\theta\\ &= \sum_{i=1}^m p_i,\end{aligned}$$

that is, the sum of the weights must be one.

We can calculate the mean and variance of a mixture distribution as follows. We will assume, for simplicity, that θ is a scalar. Let $E_i(\theta)$ and $Var_i(\theta)$ be the mean and variance of the distribution for θ in component i, that is,

$$E_i(\theta) = \int_\Theta \theta\,\pi_i(\theta)\,d\theta \quad \text{and} \quad Var_i(\theta) = \int_\Theta \{\theta - E_i(\theta)\}^2\,\pi_i(\theta)\,d\theta.$$

It can be shown that the mean of the mixture distribution is

$$E(\theta) = \sum_{i=1}^m p_i E_i(\theta). \tag{1.12}$$

We also have

$$\begin{aligned} E(\theta^2) &= \sum_{i=1}^{m} p_i E_i(\theta^2) \\ &= \sum_{i=1}^{m} p_i \left\{Var_i(\theta) + E_i(\theta)^2\right\} \end{aligned} \tag{1.13}$$

from which we can calculate the variance of the mixture distribution using

$$Var(\theta) = E(\theta^2) - E(\theta)^2.$$

Combining a mixture prior distribution with data $\boldsymbol{x}$ using Bayes Theorem produces the posterior density

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &= \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})} \\ &= \sum_{i=1}^{m} \frac{p_i \pi_i(\theta)\, f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})} \end{aligned} \tag{1.14}$$

where $f(\boldsymbol{x})$ is a constant with respect to θ. Now if the prior density were $\pi_i(\theta)$ (instead of the mixture distribution), using Bayes Theorem, the posterior density would be

$$\pi_i(\theta|\boldsymbol{x}) = \frac{\pi_i(\theta)\, f(\boldsymbol{x}|\theta)}{f_i(\boldsymbol{x})}$$

where $f_i(\boldsymbol{x})$, $i = 1, 2, \ldots, m$ are constants with respect to θ, that is $\pi_i(\theta)\, f(\boldsymbol{x}|\theta) = f_i(\boldsymbol{x})\, \pi_i(\theta|\boldsymbol{x})$. Substituting this in to (1.14) gives

$$\pi(\theta|\boldsymbol{x}) = \sum_{i=1}^{m} \frac{p_i f_i(\boldsymbol{x})}{f(\boldsymbol{x})}\, \pi_i(\theta|\boldsymbol{x}).$$

Thus the posterior distribution is a mixture distribution of component distributions $\pi_i(\theta|\boldsymbol{x})$ with weights $p_i^* = p_i f_i(\boldsymbol{x})/f(\boldsymbol{x})$. Now

$$\sum_{i=1}^{m} p_i^* = 1 \quad \Rightarrow \quad \sum_{i=1}^{m} \frac{p_i f_i(\boldsymbol{x})}{f(\boldsymbol{x})} = 1 \quad \Rightarrow \quad f(\boldsymbol{x}) = \sum_{i=1}^{m} p_i f_i(\boldsymbol{x})$$

and so

$$p_i^* = \frac{p_i f_i(\boldsymbol{x})}{\displaystyle\sum_{j=1}^{m} p_j f_j(\boldsymbol{x})}, \qquad i = 1, 2, \ldots, m.$$

Hence, combining data $\boldsymbol{x}$ with a mixture prior distribution $(p_i, \pi_i(\theta))$ produces a posterior mixture distribution $(p_i^*, \pi_i(\theta|\boldsymbol{x}))$. The effect of introducing the data is to "update" the mixture weights ($p_i \rightarrow p_i^*$) and the component distributions ($\pi_i(\theta) \rightarrow \pi_i(\theta|\boldsymbol{x})$).

Example 1.13

Suppose we have a random sample of size 20 from an exponential distribution, that is, $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, 20$ (independent). Also suppose that the prior distribution for θ is the mixture distribution

$$\theta \sim 0.6\,Ga(5, 10) + 0.4\,Ga(15, 10),$$

as shown in Figure 1.7. Here the component distributions are $\pi_1(\theta) = Ga(5, 10)$ and $\pi_2(\theta) = Ga(15, 10)$, with weights $p_1 = 0.6$ and $p_2 = 0.4$.

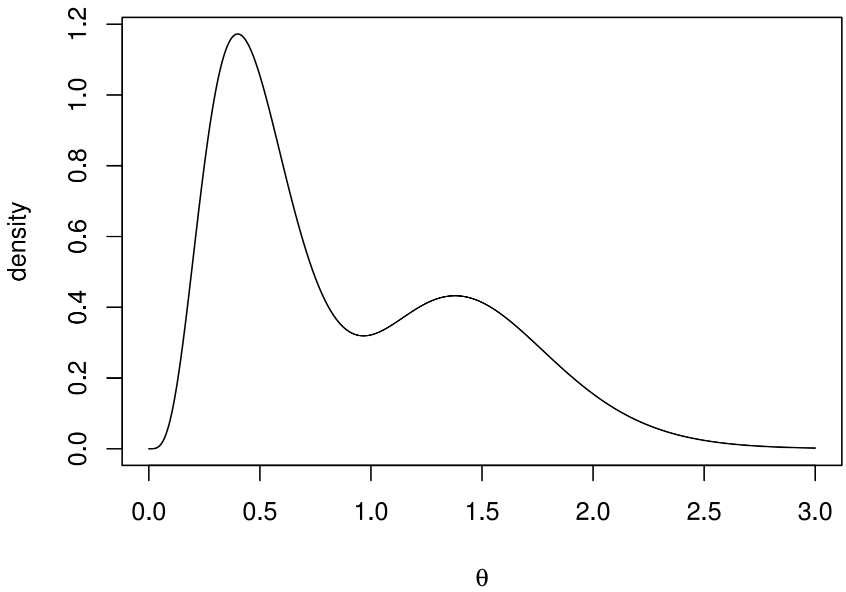

Figure 1.7: Plot of the mixture prior density

Using (1.12), the prior mean is

$$E(\theta) = 0.6 \times \frac{5}{10} + 0.4 \times \frac{15}{10} = 0.9$$

and, using (1.13), the prior second moment for θ is

$$E(\theta^2) = 0.6 \times \frac{5 \times 6}{10^2} + 0.4 \times \frac{15 \times 16}{10^2} = 1.14$$

from which we calculate the prior variance as

$$Var(\theta) = E(\theta^2) - E(\theta)^2 = 1.14 - 0.9^2 = 0.33$$

and prior standard deviation as

$$SD(\theta) = \sqrt{Var(\theta)} = \sqrt{0.33} = 0.574.$$

We have already seen that combining a random sample of size 20 from an exponential distribution with a $Ga(g, h)$ prior distribution results in a $Ga(g + 20, h + 20\bar{x})$ posterior distribution. Therefore, the (overall) posterior distribution will be a mixture distribution with component distributions

$$\pi_1(\theta|\boldsymbol{x}) = Ga(25, 10 + 20\bar{x}) \quad \text{and} \quad \pi_2(\theta|\boldsymbol{x}) = Ga(35, 10 + 20\bar{x}).$$

We now calculate new values for the weights p_1^* and $p_2^* = 1 - p_1^*$, which will depend on both prior information and the data. We have

$$p_1^* = \frac{0.6 f_1(\boldsymbol{x})}{0.6 f_1(\boldsymbol{x}) + 0.4 f_2(\boldsymbol{x})}$$

from which

$$\left(p_1^*\right)^{-1} - 1 = \frac{0.4 f_2(\boldsymbol{x})}{0.6 f_1(\boldsymbol{x})}.$$

In general, the functions

$$f_i(\boldsymbol{x}) = \int_\Theta \pi_i(\theta)\, f(\boldsymbol{x}|\theta)\, d\theta$$

are potentially complicated integrals (solved either analytically or numerically). However, as with Candidates formula, these calculations become much simpler when we have a conjugate prior distribution: rewriting Bayes Theorem, we obtain

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

and so when the prior and posterior densities have a simple form (as they do when using a conjugate prior), it is straightforward to determine $f(\boldsymbol{x})$ using algebra rather than having to use calculus.

In this example we know that the gamma distribution is the conjugate prior distribution: using a random sample of size n with mean $\bar{x}$ and a $Ga(g, h)$ prior distribution gives a $Ga(g + n, h + n\bar{x})$ posterior distribution, and so

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{\pi(\theta|\boldsymbol{x})} \\ &= \frac{\dfrac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \theta^n e^{-n\bar{x}\theta}}{\dfrac{(h + n\bar{x})^{g+n} \theta^{g+n-1} e^{-(h+n\bar{x})\theta}}{\Gamma(g+n)}} \\ &= \frac{h^g\, \Gamma(g+n)}{\Gamma(g)(h + n\bar{x})^{g+n}}. \end{aligned}$$

Therefore

$$\begin{aligned} \left(p_1^*\right)^{-1} - 1 &= \frac{0.4 \times 10^{15}\, \Gamma(35)}{\Gamma(15)(10 + 20\bar{x})^{35}} \Bigg/ \frac{0.6 \times 10^{5}\, \Gamma(25)}{\Gamma(5)(10 + 20\bar{x})^{25}} \\ &= \frac{2\Gamma(35)\Gamma(5)}{3\Gamma(25)\Gamma(15)(1 + 2\bar{x})^{10}} \\ &= \frac{611320}{7(1 + 2\bar{x})^{10}} \end{aligned}$$

and so

$$p_1^* = \frac{1}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}}, \qquad p_2^* = 1 - p_1^*.$$

Hence the posterior distribution is the mixture distribution

$$\frac{1}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}} \times Ga(25, 10+20\bar{x}) + \left(1 - \frac{1}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}}\right) \times Ga(35, 10+20\bar{x}).$$

Recall that the most likely value of θ from the data alone, the likelihood mode, is $1/\bar{x}$. Therefore, large values of $\bar{x}$ indicate that θ is small and *vice versa*. With this in mind, it is not surprising that the weight p_1^* (of the component distribution with the smallest mean) is increasing in $\bar{x}$, and $p_1^* \to 1$ as $\bar{x} \to \infty$. Using (1.12), the posterior mean is

$$\begin{aligned}E(\theta|\boldsymbol{x}) &= \frac{1}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}} \times \frac{25}{10+20\bar{x}} + \left(1 - \frac{1}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}}\right) \times \frac{35}{10+20\bar{x}} \\ &= \cdots \\ &= \frac{1}{2(1+2\bar{x})}\left\{7 - \frac{2}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}}\right\}.\end{aligned}$$

The posterior standard deviation can be calculated using (1.12) and (1.13).

Table 1.6 shows the posterior distributions which result when various sample means $\bar{x}$ are observed together with the posterior mean and the posterior standard deviation. Graphs of these posterior distributions, together with the prior distribution, are given in Figure 1.8. When considering the effect on beliefs of observing the sample mean $\bar{x}$, it is important to remember that large values of $\bar{x}$ indicate that θ is small and *vice versa*. Plots of the posterior mean against the sample mean reveal that the posterior mean lies between the prior mean and the likelihood mode only for $\bar{x} \in (0, 0.70) \cup (1.12, \infty)$. Note that observing the data has focussed our beliefs about θ in the sense that the posterior standard deviation is less than the prior standard deviation – and considerably less in some cases.

$\bar{x}$	$\hat{\theta} = 1/\bar{x}$	Posterior mixture distribution	$E(\theta\|\boldsymbol{x})$	$SD(\theta\|\boldsymbol{x})$
4	0.25	$0.99997\,Ga(25,90) + 0.00003\,Ga(35,90)$	0.278	0.056
2	0.5	$0.9911\,Ga(25,50) + 0.0089\,Ga(35,50)$	0.502	0.102
1.2	0.8	$0.7027\,Ga(25,34) + 0.2973\,Ga(35,34)$	0.823	0.206
1	1.0	$0.4034\,Ga(25,30) + 0.5966\,Ga(35,30)$	1.032	0.247
0.8	1.25	$0.1392\,Ga(25,26) + 0.8608\,Ga(35,26)$	1.293	0.260
0.5	2.0	$0.0116\,Ga(25,20) + 0.9884\,Ga(35,20)$	1.744	0.300

Table 1.6: Posterior distributions (with summaries) for various sample means $\bar{x}$

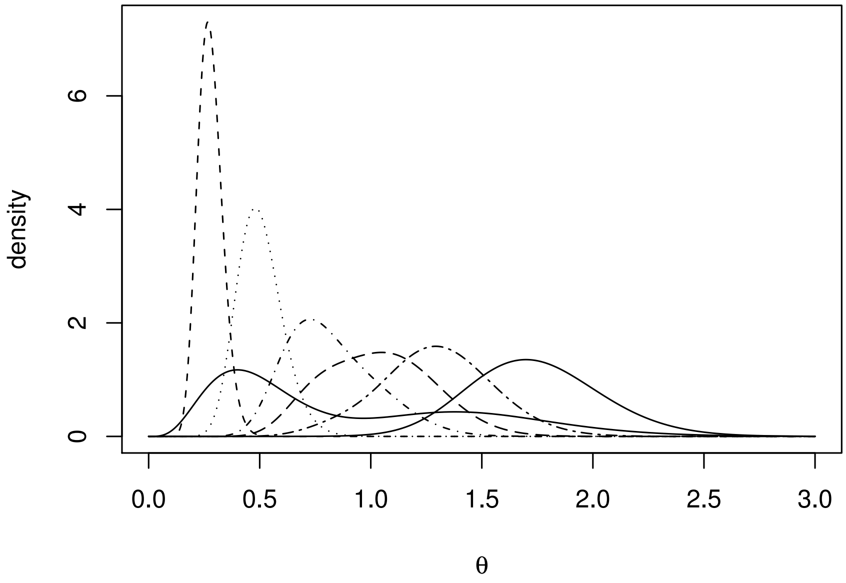

Figure 1.8: Plot of the prior distribution and various posterior distributions

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
- determine posterior distributions when the prior is a mixture of conjugate distributions