Chapter 2

Bayes' Theorem for Distributions

2.1 Introduction

Suppose we have data $\boldsymbol{x}$ 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.

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

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

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

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

Thus, to obtain the posterior distribution, we need:

(1) data, from which we can form the **likelihood** $f(\boldsymbol{x}|\theta)$, and

(2) a suitable distribution, $\pi(\theta)$, that represents our **prior beliefs** about θ.

You should now be comfortable with how to obtain the likelihood (point 1 above; see Section 1.3 of these notes, plus MAS2901!). But how do we specify a prior (point 2)? In Chapter 3 we will consider the task of *prior elicitation* – the process which facilitates the 'derivation' of a suitable prior distribution for θ. For now, we will assume someone else has done this for us; the main aim of this chapter is simply to operate Bayes' Theorem for distributions to obtain the posterior distribution for θ. And before we do this, it will be worth re–familiarising ourselves with some continuous probability distributions you have met before, and which we will use extensively in this course: the uniform, beta and gamma distributions (indeed, I will assume that you are more than familiar with some other 'standard' distributions we will use – e.g. the exponential, Normal, Poisson, and binomial, and so will *not* review these here).

Definition 2.1 (Continuous Uniform distribution)

The random variable Y follows a Uniform $U(a, b)$ distribution if it has probability density function

$$f(y|a, b) = \frac{1}{b-a}, \qquad a \leq y \leq b.$$

This form of probability density function ensures that all values in the range $[a, b]$ are *equally likely*, hence the name "uniform". This distribution is sometimes called the *rectangular distribution* because of its shape.

You should remember from MAS1604 that

$$E(Y) = \frac{a+b}{2} \quad \text{and} \quad Var(Y) = \frac{(b-a)^2}{12}.$$

In the space below, sketch the probability density functions for $U(0, 1)$ and $U(10, 50)$.

Definition 2.2 (Beta distribution)

The random variable Y follows a Beta $Be(a, b)$ distribution ($a > 0$, $b > 0$) if it has probability density function

$$f(y|a, b) = \frac{y^{a-1}(1-y)^{b-1}}{B(a, b)}, \qquad 0 < y < 1. \tag{2.1}$$

The constant term $B(a, b)$, also known as the *beta function*, ensures that the density integrates to one. Therefore

$$B(a, b) = \int_0^1 y^{a-1}(1-y)^{b-1}\, dy. \tag{2.2}$$

It can be shown that the beta function can be expressed in terms of another function, called the *gamma function* $\Gamma(\cdot)$, as

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\, dx. \tag{2.3}$$

Tables are available for both $B(a, b)$ and $\Gamma(a)$. However, these functions are very simple to evaluate when a and b are integers since the gamma function is a generalisation of the factorial function. In particular, when a and b are integers, we have

$$\Gamma(a) = (a-1)! \quad \text{and} \quad B(a, b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}.$$

For example,

$$B(2, 3) = \frac{1! \times 2!}{4!} = \frac{1}{12}.$$

It can be shown, using the identity $\Gamma(a) = (a-1)\Gamma(a-1)$, that

$$E(Y) = \frac{a}{a+b}, \qquad \text{and} \qquad Var(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Also

$$Mode(Y) = \frac{a-1}{a+b-2}, \quad \text{if } a > 1 \text{ and } b > 1.$$

Definition 2.3 (Gamma distribution)

The random variable Y follows a Gamma $Ga(a, b)$ distribution ($a > 0$, $b > 0$) if it has probability density function

$$f(y|a, b) = \frac{b^a y^{a-1} e^{-by}}{\Gamma(a)}, \qquad y > 0,$$

where $\Gamma(a)$ is the gamma function defined in (2.3). It can be shown that

$$E(Y) = \frac{a}{b} \qquad \text{and} \qquad Var(Y) = \frac{a}{b^2}.$$

Also

$$Mode(Y) = \frac{a-1}{b}, \quad \text{if } a \geq 1.$$

We can use R to visualise the beta and gamma distributions for various values of (a, b) (and indeed any other standard probability distribution you have met so far). For example, we know that the beta distribution is valid for all values in the range $(0, 1)$. In R, we can set this up by typing:

```
> x=seq(0,1,0.01)
```

which specifies x to take all values in the range 0 to 1, in steps of 0.01. The following code then calculates the density of $Be(2, 5)$, as given by Equation (2.1) with $a = 2$ and $b = 5$:

```
> y=dbeta(x,2,5)
```

Plotting y against x and joining with lines gives the $Be(2, 5)$ density shown in Figure 2.1 (top left); in R this is achieved by typing

```
> plot(x,y,type='l')
```

Also shown in Figure 2.1 are densities for the $Be(0.5, 0.5)$ (top right), $Be(77, 5)$ (bottom left) and $Be(10, 10)$ (bottom right) distributions. Notice that different combinations of (a, b) give rise to different shapes of distribution between the limits of 0 and 1 – symmetric, positively skewed and negatively skewed: careful choices of a and b could thus be used to express our prior beliefs about probabilities/proportions we think might be more or less likely to occur. When $a = b$ we have a distribution which is symmetric about 0.5. Similar plots can be constructed for any standard distribution of interest using, for example, `dgamma` or `dnorm` instead of `dbeta` for the gamma or Normal distributions, respectively; Figure 2.2 shows densities for various gamma distributions.

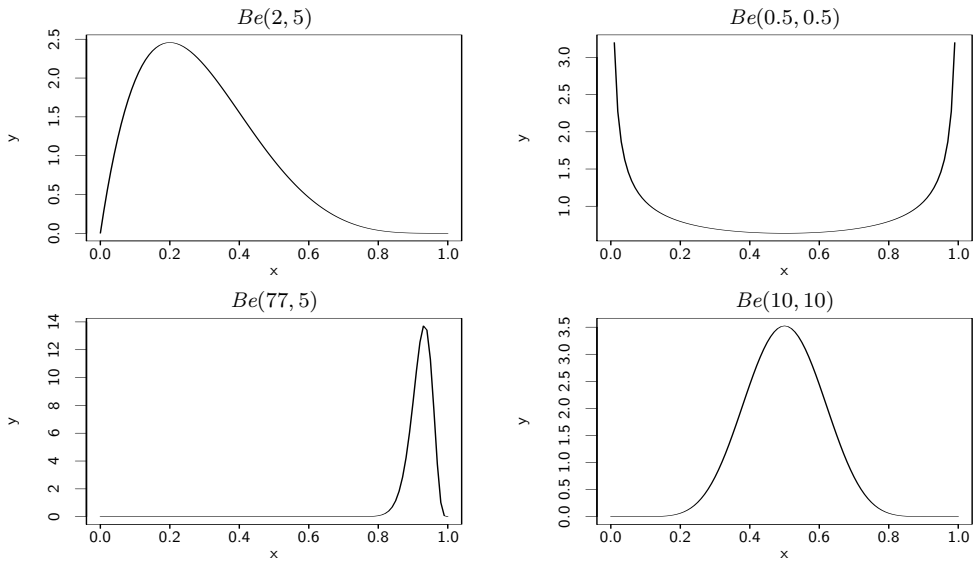

Figure 2.1: Plots of $Be(a,b)$ densities for various values of (a,b).

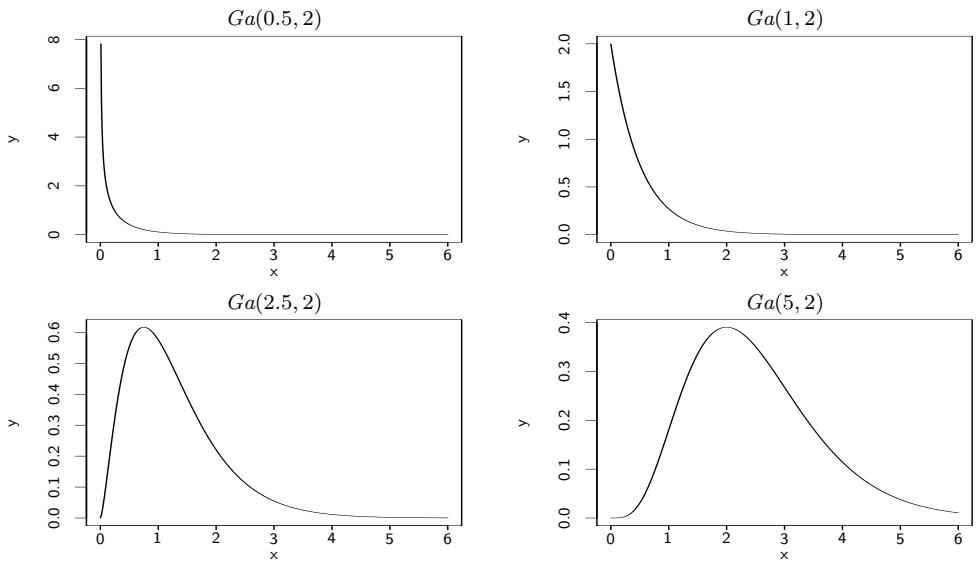

Figure 2.2: Plots of $Ga(a, b = 2)$ densities, for various values of a.

2.2 Bayes' Theorem for distributions in action

We will now see Bayes' Theorem for distributions in operation. Remember – for now, we will assume that someone else has derived the prior distribution for θ for us. In Chapter 3 we will consider how this might be done.

Example 2.1

Consider an experiment with a possibly biased coin. Let $\theta = \Pr(\text{Head})$. Suppose that, before conducting the experiment, we believe that all values of θ are equally likely: this gives a prior distribution $\theta \sim U(0, 1)$, and so

$$\pi(\theta) = 1, \qquad 0 < \theta < 1. \tag{2.4}$$

Note that with this prior distribution $E(\theta) = 0.5$. We now toss the coin 5 times and observe 1 head. Determine the posterior distribution for θ given this data.

✎ *...Solution to Example 2.1...*

✎ *...Solution to Example 2.1 continued...*

The main difficulty in calculating the posterior distribution was in obtaining the $f(x)$ term (2.6). However, in many cases we can recognise the posterior distribution without the need to calculate this constant term (constant with respect to θ). In this example, we can calculate the posterior distribution as

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)f(x=1|\theta)\\ &\propto 1\times 5\theta(1-\theta)^4, \qquad 0<\theta<1\\ &= k\theta(1-\theta)^4, \qquad 0<\theta<1.\end{aligned}$$

As θ is a continuous quantity, what we would like to know is what continuous distribution defined on $(0,1)$ has a probability density function which takes the form $k\theta^{g-1}(1-\theta)^{h-1}$. The answer is the $Be(g,h)$ distribution. Therefore, choosing g and h appropriately, we can see that the posterior distribution is $\theta|x=1 \sim Be(2,5)$.

Summary:

It is possible that we have a biased coin. If we suppose that all values of $\theta = \Pr(\text{Head})$ are equally likely and then observe 1 head out of 5, then the most likely value of θ is 0.2 — the same as the most likely value from the data alone (not surprising!). However, on average, we would expect θ to be around 0.286. Uncertainty about θ has changed from a (prior) standard deviation of 0.289 to a (posterior) standard deviation of 0.160. The changes in our beliefs about θ are more fully described by the prior and posterior distributions shown in Figure 2.3.

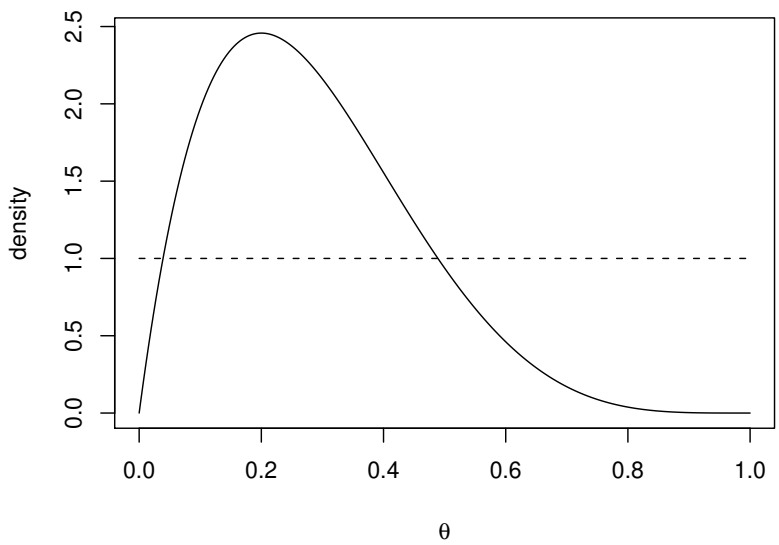

Figure 2.3: Prior (dashed) and posterior (solid) densities for $\theta = \Pr(\text{Head})$

Example 2.2

Consider an experiment to determine how good a music expert is at distinguishing between pages from Haydn and Mozart scores. Let $\theta = \Pr(\text{correct choice})$. Suppose that, before conducting the experiment, we have been told that the expert is very competent. In fact, it is suggested that we should have a prior distribution which has a mode around $\theta = 0.95$ and for which $\Pr(\theta < 0.8)$ is very small. We choose $\theta \sim Be(77, 5)$ (see Example 3.2, Chapter 3), with probability density function

$$\pi(\theta) = 128107980\,\theta^{76}(1-\theta)^4, \qquad 0 < \theta < 1. \tag{2.7}$$

A graph of this prior density is given in Figure 2.4.

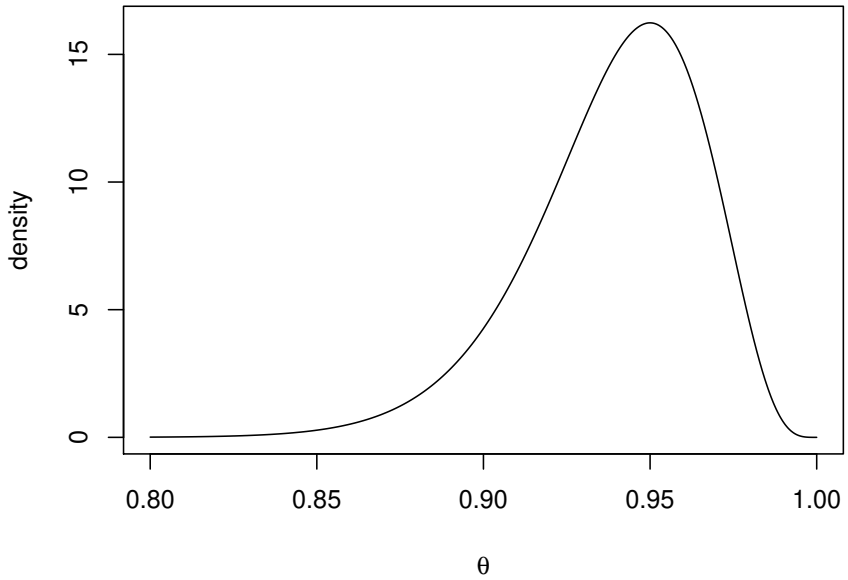

Figure 2.4: Prior density for the music expert's skill.

In the experiment, the music expert makes the correct choice 9 out of 10 times. Determine the posterior distribution for θ given this information.

...Solution to Example 2.2...

✎ *...Solution to Example 2.2 continued...*

Summary:

The changes in our beliefs about θ are described by the prior and posterior distributions shown in Figure 2.5 and summarised in Table 2.1.

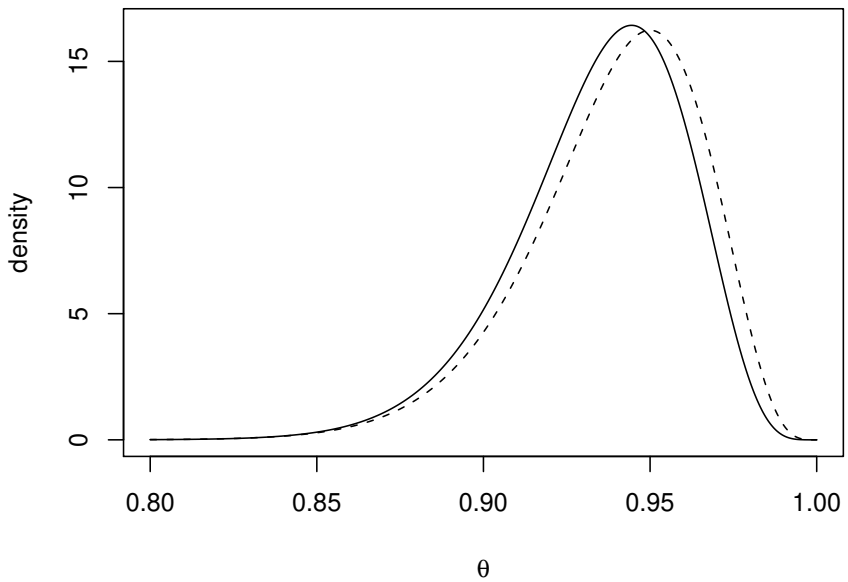

Figure 2.5: Prior (dashed) and posterior (solid) densities for the music expert's skill.

	Prior (2.7)	Likelihood (2.8)	Posterior (2.9)
$Mode(\theta)$	0.950	0.900	0.944
$E(\theta)$	0.939	–	0.935
$SD(\theta)$	0.0263	–	0.0256

Table 2.1: Changes in beliefs about θ.

Notice that, having observed only a 90% success rate in the experiment, the posterior mode and mean are smaller than their prior values. Also, the experiment has largely confirmed our ideas about θ, with the uncertainty about θ being only very slightly reduced.

Example 2.3

Max, a video game pirate, is trying to identify the proportion of potential customers θ who might be interested in buying *Call of Duty: Elite* next month. Based on the proportion of customers who have bought similarly violent games from him in the past, he assumes that $\theta \sim Be(2.5, 12)$ (see Example 3.3, Chapter 3); a plot of this prior density is shown in Figure 2.6.

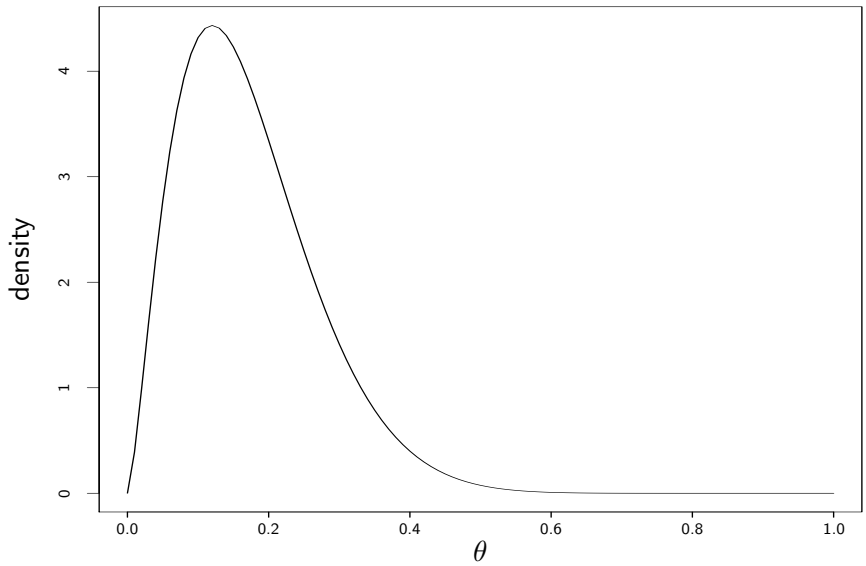

Figure 2.6: Max's prior density.

Max asks five potential customers if they would buy *Call of Duty: Elite* from him, and four say they would. Using this information, what is Max's posterior distribution for θ?

...Solution to Example 2.3...

✎ *...Solution to Example 2.3 continued...*

Summary:

The changes in our beliefs about θ are described by the prior and posterior distributions shown in Figure 2.7 and summarised in Table 2.2.

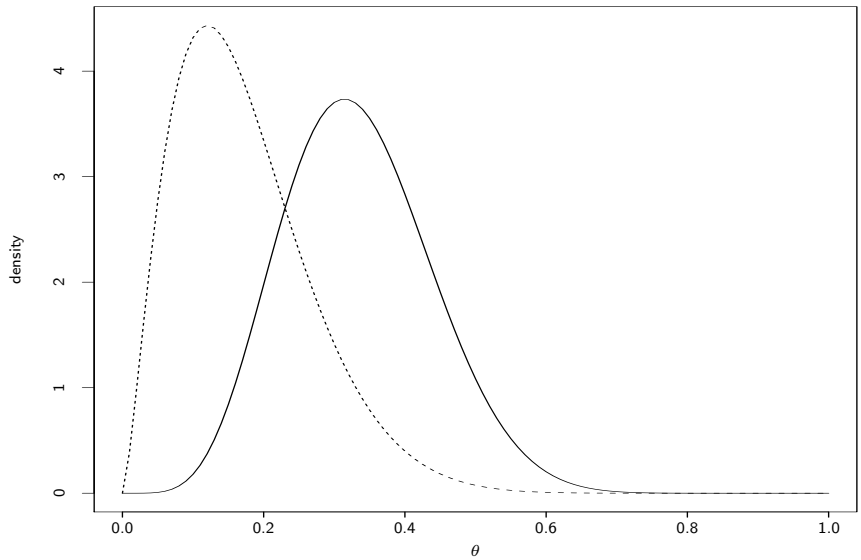

Figure 2.7: Prior (dashed) and posterior (solid) densities for Max's problem.

	Prior (2.10)	Likelihood (2.11)	Posterior (2.12)
$Mode(\theta)$	0.12	0.8	0.314
$E(\theta)$	0.172	–	0.333
$SD(\theta)$	0.096	–	0.104

Table 2.2: Changes in beliefs about θ.

Notice how the posterior has been "pulled" from the prior towards the observed value: the mode has moved up from 0.12 to 0.314, and the mean has moved up from 0.172 to 0.333. Having just one observation in the likelihood, we see that there is hardly any change in the standard deviation from prior to posterior: we would expect to see a decrease in standard deviation with the addition of more data values.

Example 2.4

Table 2.3 shows some data on the times between serious earthquakes. An earthquake is included if its magnitude is at least 7.5 on the Richter scale or if over 1000 people were killed. Recording starts on 16 December 1902 (4500 killed in Turkistan). The table includes data on 21 earthquakes, that is, 20 "waiting times" between earthquakes.

840	157	145	44	33	121	150	280	434	736
584	887	263	1901	695	294	562	721	76	710

Table 2.3: Time intervals between major earthquakes (in days).

It is believed that earthquakes happen in a random haphazard kind of way and that times between earthquakes can be described by an exponential distribution. Data over a much longer period suggest that this exponential assumption is plausible. Therefore, we will assume that these data are a random sample from an exponential distribution with rate θ (and mean $1/\theta$). The parameter θ describes the rate at which earthquakes occur.

An expert on earthquakes has prior beliefs about the rate of earthquakes, θ, described by a $Ga(10, 4000)$ distribution (see Example 3.1, Chapter 3), which has density density

$$\pi(\theta) = \frac{4000^{10}\,\theta^9 e^{-4000\theta}}{\Gamma(10)}, \quad \theta > 0, \tag{2.13}$$

and mean $E(\theta) = 0.0025$. A plot of this prior distribution can be found in Figure 2.8. As you might expect, the expert believes that, realistically, only very small values of θ are likely, though larger values are not ruled out! Determine the posterior distribution for θ.

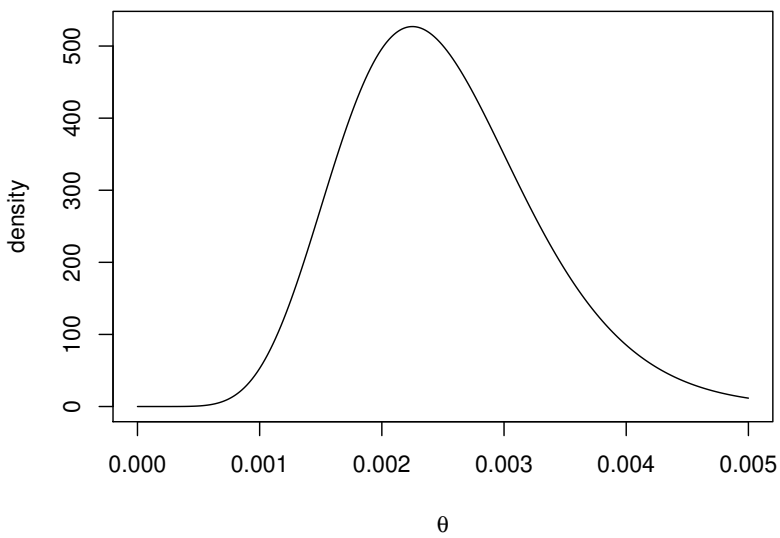

Figure 2.8: Prior density for the earthquake rate θ

✎ *...Solution to Example 2.4...*

Summary:

The data have updated our beliefs about θ from a $Ga(10, 4000)$ distribution to a $Ga(30, 13633)$ distribution. Plots of these distributions are given in Figure 2.9, and Table 2.4 gives a summary of the main changes induced by incorporating the data. 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. This results in 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 likely earthquake rate θ.

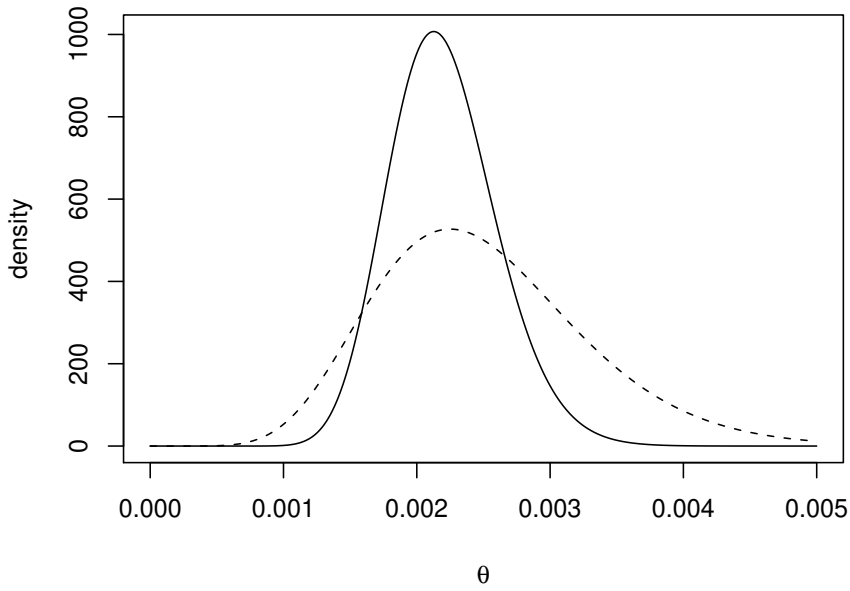

Figure 2.9: Prior (dashed) and posterior (solid) densities for the earthquake rate θ.

	Prior (2.13)	Likelihood (2.14)	Posterior (2.15)
$Mode(\theta)$	0.00225	0.00208	0.00213
$E(\theta)$	0.00250	–	0.00220
$SD(\theta)$	0.00079	–	0.00040

Table 2.4: Changes in beliefs about θ.

Example 2.5

We now consider the general case of the problem discussed in Example 2.4. Suppose $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (independent) and our prior beliefs about θ are summarised by a $Ga(g, h)$ distribution (with g and h known), with density

$$\pi(\theta) = \frac{h^g \, \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \tag{2.16}$$

Determine the posterior distribution for θ.

✎ *...Solution to Example 2.5...*

Summary:

If we have a random sample from an $Exp(\theta)$ 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 $Ga(g + n, h + n\bar{x})$ distribution.

The changes in our beliefs about θ are summarised in Table 2.5, taking $g \geq 1$.

	Prior (2.16)	Likelihood (2.17)	Posterior (2.18)
$Mode(\theta)$	$(g-1)/h$	$1/\bar{x}$	$(g+n-1)/(h+n\bar{x})$
$E(\theta)$	g/h	–	$(g+n)/(h+n\bar{x})$
$SD(\theta)$	$\sqrt{g}/h$	–	$\sqrt{g+n}/(h+n\bar{x})$

Table 2.5: Changes in beliefs about θ.

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,

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

The standard deviation of the posterior distribution is smaller than that of the prior distribution if and only if the sample mean is large enough, that is

$$SD(\theta|\boldsymbol{x}) < SD(\theta) \quad \Longleftrightarrow \quad \bar{x} > k.$$

Conjugacy

In Example 2.5, a gamma prior leads to a gamma posterior. This is an example of **conjugacy**, where choosing a prior in a family of distributions always leads to a posterior in the same family. The formal definition is as follows.

Definition 2.4

Suppose that data $\boldsymbol{x}$ are modelled 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 $\boldsymbol{x}$. So we say (for example) that the Gamma distribution is a conjugate prior for the exponential model.

There are usually simple formulae to update the conjugate prior to the corresponding posterior. For instance, Example 2.5 showed that a $Ga(g, h)$ prior for an exponential model leads to a $Ga(g + n, h + n\bar{x})$ distribution. For this reason, conjugate priors are

usually a convenient choice to reduce the mathematical and computational effort of deriving the posterior.

Some examples of conjugacy are as follows:

1. Binomial random sample, Beta prior distribution $\longrightarrow$ Beta posterior distribution (Examples 2.1 – 2.3)
2. Exponential random sample, Gamma prior distribution $\longrightarrow$ Gamma posterior distribution (Examples 2.4 – 2.5)
3. Normal random sample (known variance), Normal prior distribution $\longrightarrow$ Normal posterior distribution (Example 2.6)
4. $Ga(k, \theta)$ random sample (k known), Gamma prior distribution $\longrightarrow$ Gamma posterior distribution.

Example 2.6

Suppose we have a random sample from a normal distribution. In Bayesian statistics, when dealing with the normal distribution, the mathematics is more straightforward if we work 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

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}. \tag{2.19}$$

Determine the posterior distribution for μ.

...Solution to Example 2.6...

✎ *...Solution to Example 2.6 continued...*

✎ *...Solution to Example 2.6 continued...*

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.

Notice that the way prior information and observed data combine is through the parameters of the normal distribution:

$$b \to \frac{db + n\tau\bar{x}}{d + n\tau} \qquad \text{and} \qquad d^2 \to d + n\tau.$$

Notice also that the posterior variance (and precision) does not depend on the data, and the posterior mean is a convex combination of the prior and sample means, that is,

$$B = \alpha b + (1 - \alpha)\bar{x},$$

for some $\alpha \in (0, 1)$. This equation for the posterior mean, which can be rewritten as

$$E(\mu|\boldsymbol{x}) = \alpha E(\mu) + (1 - \alpha)\bar{x},$$

arises in other models and is known as the *Bayes linear rule*.

The changes in our beliefs about μ are summarised in Table 2.6. 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

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

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

	Prior (2.19)	Likelihood (2.20)	Posterior (2.22)
$Mode(\mu)$	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 2.6: Changes in beliefs about μ.

Example 2.7

The ages of *Ennerdale granophyre* rocks can be determined using the relative proportions of rubidium–87 and strontium–87 in the rock. An expert in the field suggests that the ages of such rocks (in millions of years) $X|\mu \sim N(\mu, 8^2)$ and that a prior distribution $\mu \sim N(370, 20^2)$ is appropriate. A rock is found whose chemical analysis yields $x = 421$. What is the posterior distribution for μ and what is the probability that the rock will be older than 400 million years?

✎ *...Solution to Example 2.7...*

This highlights the benefit of taking the chemical measurements. Note that the large difference between these probabilities is not necessarily due to the expert's prior distribution being inaccurate, *per se*, it is probably due to the large prior uncertainty about rock ages, as shown in Figure 2.10.

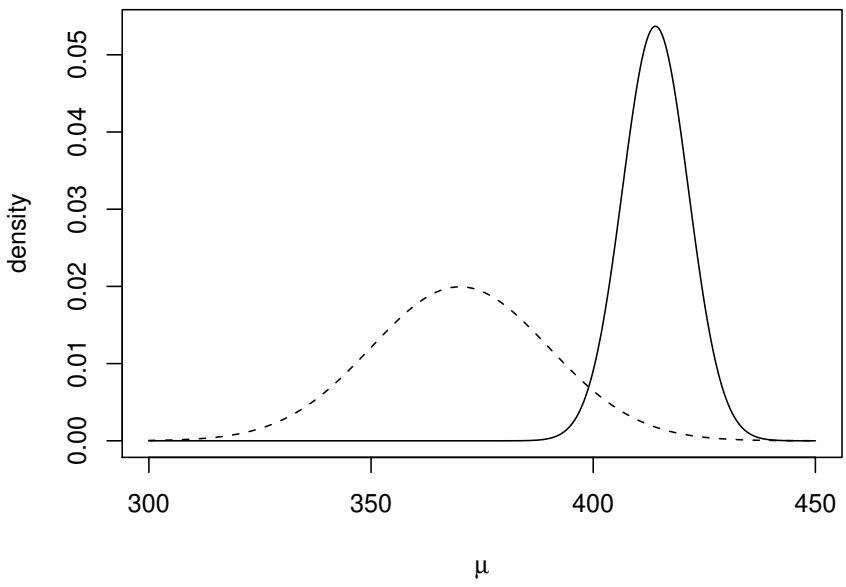

Figure 2.10: Prior (dashed) and posterior (solid) densities for the age of the rock

2.3 Posterior distributions and sufficient statistics

We have already met the concept of minimal sufficient statistics. Not surprisingly they also play a role in Bayesian Inference.

Suppose that we have data $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ available and we want to make inferences about the parameters θ in the statistical model $f(\boldsymbol{x}|\theta)$. If T is a minimal sufficient statistic then by the Factorisation Theorem

$$f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\, g(t, \theta)$$

for some functions h and g. Therefore, using Bayes Theorem

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta)\\ &\propto \pi(\theta)\, h(\boldsymbol{x})\, g(t, \theta)\\ &\propto \pi(\theta)\, g(t, \theta).\end{aligned}$$

Now it can be shown that, up to a constant not depending on θ, $g(t, \theta)$ is equal to the probability (density) function of T, that is,

$$g(t, \theta) \propto f_T(t|\theta).$$

Hence

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta)\, f_T(t|\theta).$$

However, applying Bayes Theorem to the data t gives

$$\pi(\theta|t) \propto \pi(\theta)\, f_T(t|\theta)$$

and so, since $\pi(\theta|\boldsymbol{x}) \propto \pi(\theta|t)$ and both are probability (density) functions, we have

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

Therefore, our (posterior) beliefs about θ having observed the full data $\boldsymbol{x}$ are the same as if we had observed only the sufficient statistic T. This is what we would expect if all the information about θ in the data were contained in the sufficient statistic.

Example 2.8

Suppose we have a random sample from an exponential distribution with a gamma prior distribution, that is, $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (independent) and $\theta \sim Ga(g, h)$. Determine a sufficient statistic T for θ and verify that $\pi(\theta|\boldsymbol{x}) = \pi(\theta|t)$.

✎ *...Solution to Example 2.8...*

✎ *...Solution to Example 2.8 continued...*

Example 2.9

Suppose we have a random sample from a normal distribution with known variance and a normal prior distribution for the mean parameter, that is, $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent) and $\mu \sim N(b, 1/d)$. Determine a sufficient statistic T for μ and verify that $\pi(\mu|\boldsymbol{x}) = \pi(\mu|t)$.

✎ *...Solution to Example 2.9...*