

Chapter 1

Introduction

Consider the following three experiments.

Experiment 1: The tea lady

The tea lady claims to know whether milk or tea is poured in first: for 10 pairs of cups of tea she makes the correct choice each time.

Experiment 2: Music expert

The expert claims he can distinguish between a page from a Haydn score and a page from a Mozart score: he does so correctly 10 times.

Experiment 3: The Drunk

A somewhat inebriated friend at a party claims they can predict the outcome of the toss of a coin: they do so correctly 10 times.

Let $\theta = \Pr(\text{correct choice})$. Let's suppose the tea lady, the music expert and the drunk *cannot* do as they claim; then, just by guessing, we could expect each of them to 'get it right' 5 times out of 10, i.e. $\theta = 1/2$. We could then test the null hypothesis $H_0 : \theta = 1/2$, giving

$$p\text{-value} = \left(\frac{1}{2}\right)^{10} = 0.00098 < 0.1\%.$$

From this p -value, we would conclude that we had very strong evidence against the null hypothesis (that the choices were made randomly), and perhaps feel justified in validating each claim. But does this make sense? Surely, we have some additional information about what values of θ are plausible for each experiment. Prior to each experiment, our beliefs about θ may be

Experiment 1: $\theta > 0.5$ – folklore (and science!) suggests this may be possible;

Experiment 2: $0.9 < \theta < 1.0$ – we expect an “expert” to be correct;

Experiment 3: $\theta = 0.5$ – no way of guessing correctly with a “fair” coin.

The traditional approach to Statistics, sometimes called *Frequentist Statistics* or *Classical Statistics*, may try to take this prior information into account by modifying the pure significance testing approach described above to an assessment of an “appropriate” hypothesis test. For example, in Experiment 2, the test may be of $H_0 : \theta \geq 0.9$ against $H_1 : \theta < 0.9$.

However, in Bayesian Statistics, we attempt to calibrate our prior information about unknown quantities by constructing a probability distribution which describes how likely we believe different values are to occur. This prior information is then combined with that from experimental data using Bayes Theorem. The key ingredients of a Bayesian analysis are

- a statistical model for the experimental data;
- quantifiable prior information about any unknown parameters.

Before we consider any detailed descriptions of Bayesian analyses, we recap the various interpretations of probability and highlight the subjective approach.

1.1 Probability

The concept of probability (chance) has been around for a very long time, particularly in the area of gambling. Games of chance have been played since about 3500 B.C.; the Egyptians started using cubical dice around 2000 B.C. The mathematical theory of probability was started around the 17th century by Galilei, Pascal and Fermat to solve (again) gambling problems. There are three main ways of understanding and thinking about probability.

Frequency interpretation

The probability of an outcome is the relative frequency with which the outcome would be obtained if the experiment were repeated a large number of times under similar conditions. For example, if a coin is tossed 1,000,000 times and a head appears n times then

$$\Pr(\text{Head}) = \frac{n}{1,000,000}.$$

We would expect this probability to be about 0.5. Most of your courses will have used the frequentist interpretation: repeated sampling ideas are fundamental to the techniques described.

Classical interpretation

This is based on the concept of equally likely outcomes resulting from ideas of symmetry. If the outcome of an experiment must be one of n different outcomes and these n outcomes are equally likely then the probability of each outcome is $1/n$.

Subjective interpretation

Your subjective probability for an outcome A represents your own judgement of the likelihood that the outcome will occur. This judgement will be based on the beliefs and information H you have at the time.

One way of determining (or quantifying) a subjective value for $\Pr_H(A)$ is to consider a series of possible bets with outcome

win $\pounds c$ if A occurs and $\pounds 0$ if A^c occurs.

How much would you be prepared to pay (stake) for placing such a bet? In terms of expected winnings, you should be prepared to stake $\pounds cp$ if you believe that $\Pr_H(A) = p$. Why?



One problem with this approach is that, in general, p will depend on c : a person who is willing to bet $\pounds 1$ on the spin of a coin to win $\pounds 2$ if it lands heads may refuse to bet if the stakes are raised to $\pounds 1000$ – most people are *risk-averse*. Therefore, we shall restrict our attention to the $c = 1$ case: pay $\pounds p$ for the bet

win $\pounds 1$ if A occurs and $\pounds 0$ if A^c occurs.

You can make sure that your bet is “honest” by randomising between whether you “host” the bet or “place” the bet. For example, suppose you believe that $\Pr_H(A) = 0.5$. An “honest” bet would mean that you would buy the bet for a maximum stake of $\pounds 0.50$. However, if you weren’t honest you might try to buy the bet for any amount less than $\pounds 0.50$, say $\pounds 0.20$. If you were hosting the bet, you would take the bet for any amount more than $\pounds p$, say for $\pounds 0.80$. These conflicting interests can be offset if, when choosing p , it is equally likely that you are hosting the bet or placing the bet. In such circumstances it is in your own interests to give the value of p that you believe to be “correct”.

Example 1.1

1. The probability $\Pr(\text{Newcastle Utd win the Championship this season})$ could only be determined using a subjective assessment.
2. The probability $\Pr(\text{M\&S student chosen at random was born in January})$ could be determined using a frequency or classical interpretation (with list of all M\&S students and their birth dates) or a subjective interpretation.
3. The probability $\Pr(\text{England win the toss at a given Test Match})$ could be determined using either the classical or subjective interpretation.

There are potential drawbacks with each of these ways of understanding probability:

Frequency interpretation

1. It does not say how many times the experiment should be repeated.
2. “Similar conditions” is a vague concept.
3. It is not appropriate for many probability calculations of one-off events.
4. Standard statistical methods using the frequentist approach are not totally objective since they require subjective judgements about the validity of probability models, choice of hypotheses and interpretation of results (for instance, see BMI example in Section 2.4 of the preface to these lecture notes).

Classical interpretation

1. Only applies to equally likely outcomes.
2. Depends on a subjective assessment of whether symmetry arguments apply.
3. It is not appropriate for many probability calculations of one-off events.

Subjective interpretation

1. It is not objective – but perhaps it is more obvious (honest) about when subjective beliefs are used.
2. It requires people to be *coherent*: they will not make any wagers which they are certain to lose; also, they will not prefer to suffer a given penalty when there is the option of another penalty which is certainly smaller. Being coherent results in, *inter alia*, that

$$\begin{aligned} \Pr(A_1|H) > \Pr(A_2|H) \quad \text{and} \quad \Pr(A_2|H) > \Pr(A_3|H) \\ \implies \Pr(A_1|H) > \Pr(A_3|H). \end{aligned}$$

Each of these interpretations use quite different methods of reasoning. In this course – unlike any other course you have taken so far – we will concentrate on the subjective interpretation and describe how, if carefully used, it can be a more useful approach than the other two methods.

1.2 Bayes' Theorem

Before we state Bayes' Theorem, we need a recap of conditional probability.

Definition 1.1

Consider two events E and F , where $\Pr(F) > 0$. The *conditional probability* of E given that F has occurred is

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

Definition 1.2

The events E_1, E_2, \dots, E_n form a *partition* of the sample space \mathcal{S} if they are disjoint events ($E_i \cap E_j = \emptyset$, $i \neq j$) with $\Pr(E_i) > 0$, $i = 1, 2, \dots, n$, and $\cup_{i=1}^n E_i = \mathcal{S}$. Figure 1.1 gives a diagram of a typical partition with an additional event F .

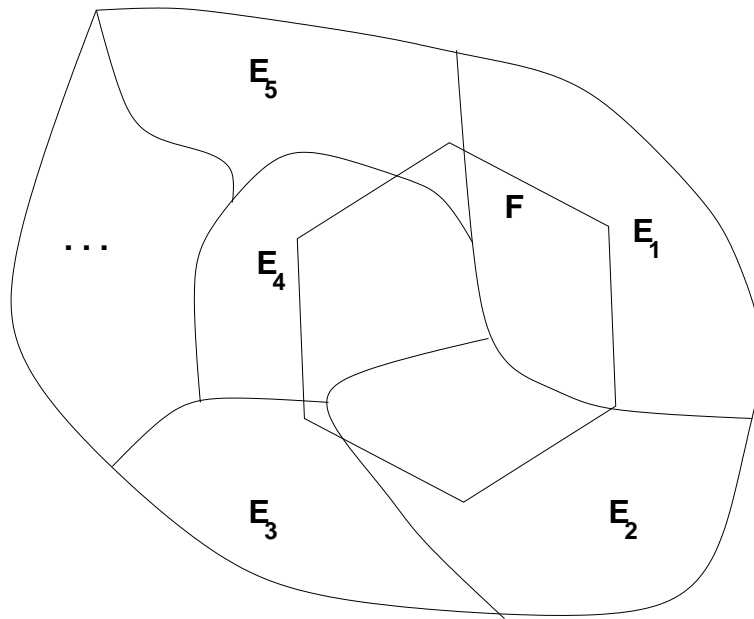


Figure 1.1: Diagram of a partition E_1, E_2, \dots, E_n and an event F

Law of Total Probability

If E_1, E_2, \dots, E_n are a *partition* of \mathcal{S} and F is any event then

$$\Pr(F) = \sum_{i=1}^n \Pr(F|E_i)\Pr(E_i).$$

Proof

As E_1, E_2, \dots, E_n are a *partition* of \mathcal{S} , we have

$$\begin{aligned}\Pr(F) &= \sum_{i=1}^n \Pr(F \cap E_i) \\ &= \sum_{i=1}^n \Pr(F|E_i)\Pr(E_i).\end{aligned}$$

Bayes' Theorem

If E_1, E_2, \dots, E_n are a *partition* of \mathcal{S} and F is any event with $\Pr(F) > 0$ then

$$\Pr(E_i|F) = \frac{\Pr(F|E_i)\Pr(E_i)}{\sum_{j=1}^n \Pr(F|E_j)\Pr(E_j)}, \quad i = 1, 2, \dots, n.$$

Proof**Example 1.2**

A laboratory blood test is 95% effective in detecting a certain disease when it is present. However, the test also yields a “false positive” result for 1% of healthy people tested. Also, 0.5% of the population actually have the disease.

- Calculate the probability that a person who tests positive actually has the disease.
- Find the probability that a person who tests negative does *not* have the disease.



...Solution to Example 1.2...

Example 1.3

Suppose that your car suffers from two intermittent problems, one caused by a fault in the engine (θ_1) and the other due to a fault in the gearbox (θ_2). These occur with probabilities 0.4 and 0.6 respectively. When examined your car exhibits one of the following symptoms

x_1 : overheating only,

x_2 : irregular traction only,

x_3 : both symptoms.

Suppose it is known in the garage trade that these symptoms occur with probabilities that depend on the fault. The probabilities $\Pr(X = x|\theta)$ are given in Table 1.1. Construct a diagnostic rule for these symptoms and determine the probability of misdiagnosis.

	O/H	I/T	Both
	x_1	x_2	x_3
θ_1 : fault in engine	0.1	0.4	0.5
θ_2 : fault in gearbox	0.5	0.3	0.2

Table 1.1: Likelihood of symptoms for both faults



...Solution to Example 1.3...

 ...Solution to Example 1.3 continued...

	O/H x_1	I/T x_2	Both x_3
θ_1 : fault in engine	0.118	0.471	0.625
θ_2 : fault in gearbox	0.882	0.529	0.375

Table 1.2: Posterior probabilities of the faults for various symptoms

This table is very informative. For example, it shows that if both symptoms (x_3) are observed, then the probability that the fault is in the engine (θ_1) changes from 0.4 to 0.625. In terms of odds

$$\text{Prior odds : } \frac{Pr(\theta_1)}{Pr(\theta_2)} = \frac{0.4}{0.6} = \frac{2}{3} \quad \text{or 3:2 in favour of } \theta_2$$

$$\text{Posterior odds : } \frac{Pr(\theta_1|x_3)}{Pr(\theta_2|x_3)} = \frac{0.625}{0.375} = \frac{5}{3} \quad \text{or 5:3 in favour of } \theta_1.$$

We are now in a position to design our diagnostic rule. This is simply a rule which diagnoses a symptom (x) as being due to some particular fault (θ). Consider first that we observe overheating only (x_1). The posterior probabilities are in favour of declaring the fault as in the gearbox (θ_2) since $Pr(\theta_2|x_1) > Pr(\theta_1|x_1)$. In the same way, we can determine the best (most likely) diagnosis having observed irregular traction only (x_2) and both symptoms (x_3), giving the diagnostic rule in Table 1.3.

Symptom	Diagnosis
overheating only (x_1)	fault in gearbox (θ_2)
irregular traction only (x_2)	fault in gearbox (θ_2)
both symptoms (x_3)	fault in engine (θ_1)


Table 1.3: Diagnostic rule for faults



...Solution to Example 1.3 continued...

Example 1.4

A student sits a multiple choice exam in which there are m alternative answers to each question. The student either knows the answer (with probability θ) or guesses randomly (with probability $1 - \theta$). What is the probability that the student actually knew the answer to a question they answered correctly?

 ...Solution to Example 1.4...

Suppose that there are $m = 5$ alternative answers for each question. We can see the effect of observing a correct answer on our belief that the student actually knows the answer by calculating $Pr(K|C)$ for various θ – see Table 1.4.

The main problem with this solution is that, in order to use this table we must know the exact value of θ : we have actually found an expression for $Pr(K|C, \theta)$ and not for $Pr(K|C)$. If we know θ fairly accurately – say it was between 0.49 and 0.51 – then, in practice, we can conclude that $Pr(K|C)$ is around 0.83. However, if we are less certain about a correct value for θ we might be able to express our uncertainty through a probability distribution for θ . We will see more about this in the next Chapter.

$Pr(K)$ $= \theta$	$Pr(K C)$ $= 5\theta/(1 + 4\theta)$
0.0	0.000
0.1	0.357
0.2	0.556
0.3	0.682
0.4	0.769
0.5	0.833
0.6	0.882
0.7	0.921
0.8	0.952
0.9	0.978
1.0	1.000

Table 1.4: Values of $Pr(K|C)$ for various values of $Pr(K)$

1.3 Likelihood

Suppose that an experiment results in data $\underline{x} = (x_1, x_2, \dots, x_n)^T$ and we decide to model the data using a probability (density) function $f_{\underline{X}}(\underline{x}|\underline{\theta})$. The likelihood function is the p(d)f of the data considered as a function of $\underline{\theta}$ for fixed \underline{x} — not a function of \underline{x} for fixed $\underline{\theta}$. We write

$$L(\underline{\theta}|\underline{x}) = f_{\underline{X}}(\underline{x}|\underline{\theta}). \quad (1.1)$$

This equation can be simplified if we have further structure in the data. For example, we may have independent observations, in which case

$$L(\underline{\theta}|\underline{x}) = f_{X_1}(x_1|\underline{\theta}) \times f_{X_2}(x_2|\underline{\theta}) \times \dots = \prod_{i=1}^n f_{X_i}(x_i|\underline{\theta}), \quad (1.2)$$

or independent and identically distributed observations (random sample), so that

$$L(\underline{\theta}|\underline{x}) = \prod_{i=1}^n f_X(x_i|\underline{\theta}). \quad (1.3)$$

In this course, we will rarely consider models with correlated observations. Moreover, we will concentrate on how to make inferences from random samples using prior information. This will require extensive use of the (1.3) form of the likelihood function.

Example 1.5

Suppose we have a random sample $\underline{x} = (x_1, x_2, \dots, x_n)^T$ of radioactive particle counts. A typical model for such data would be $X_i|\theta \sim \text{Poisson}(\theta)$, (independent). Determine the likelihood function for θ .



...Solution to Example 1.5...

Example 1.6


Suppose we have a random sample $\underline{x} = (x_1, x_2, \dots, x_n)^T$ of times between radioactive particle emissions. If the emissions occur randomly in time then a plausible model for such data would be $X_i|\theta \sim \text{Exponential}(\theta)$, (independent). Determine the likelihood function for θ .



...Solution to Example 1.6...

Example 1.7

Suppose we had observations $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ from a simple linear regression model: $Y_i|X_i = x, \alpha, \beta, \sigma \sim N(\alpha + \beta x, \sigma^2)$, (independent). Determine the likelihood function for (α, β, σ) .

 ...Solution to Example 1.7...

1.4 Sufficiency

Consider again the Poisson model in Example 1.5. The likelihood function is

$$L(\theta|\underline{x}) = \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!}$$

$$= \left(\prod_{i=1}^n x_i! \right)^{-1} \times \theta^{n\bar{x}} e^{-n\theta}.$$

Notice that this likelihood function depends on the data only through $\prod_{i=1}^n (x_i!)^{-1}$ and \bar{x} . Further, in L , θ only “interacts” with \bar{x} — the other term simply scales L — so that, for example, the point at which L is maximized is determined only by \bar{x} . Informally, we think of all the information about θ in the data being contained in \bar{x} . More formally, we can show that the distribution of the data given the value \bar{x} does not depend on θ .

Definition 1.3

A *statistic* is any function of the data (and not of unknown parameters).

Definition 1.4

The statistic $T(\underline{X})$ is *sufficient* for θ if $f_{\underline{X}}(\underline{x}|T(\underline{X}) = t)$ does not depend on θ .

Example 1.8

Consider again the Poisson model in Example 1.5. Suppose we had just two observations. Then $n = 2$ and $X_i|\theta \sim \text{Poisson}(\theta)$, $i = 1, 2$ (independent). Show that $T = X_1 + X_2$ is sufficient for θ . Note that $T|\theta \sim \text{Poisson}(2\theta)$.



...Solution to Example 1.8...

 *...Solution to Example 1.8 continued...*

Definition 1.5

The statistics $\underline{T}(\underline{X}) = (T_1(\underline{X}), T_2(\underline{X}), \dots, T_k(\underline{X}))^T$ are (jointly) *sufficient* for $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T$ if $f_{\underline{X}}(\underline{x}|\underline{T}(\underline{X}) = \underline{t})$ does not depend on $\underline{\theta}$.

Factorisation Theorem

Under certain regularity conditions

$$\underline{T} \text{ is sufficient for } \underline{\theta} \iff f_{\underline{X}}(\underline{x}|\underline{\theta}) = h(\underline{x}) g(\underline{T}, \underline{\theta}) \text{ for some functions } h \text{ and } g.$$

Example 1.9

Consider again the Poisson model in Example 1.5 with $n = 2$: $X_i|\theta \sim \text{Poisson}(\theta)$, $i = 1, 2$ (independent). Determine a sufficient statistic for θ .



...Solution to Example 1.9...

Example 1.10


Suppose we have a random sample from an exponential distribution: $X_i|\theta \sim \text{Exponential}(\theta)$, $i = 1, 2, \dots, n$ (independent). Determine a sufficient statistic for θ .



...Solution to Example 1.10...

Example 1.11

Suppose we have a random sample from a normal distribution: $X_i|\mu, \sigma \sim N(\mu, \sigma^2)$, $i = 1, 2, \dots, n$ (independent). Determine sufficient statistics for (μ, σ) .

 *...Solution to Example 1.11...*

Comment

If a parameter has a sufficient statistic then, in fact, it has an infinite number of sufficient statistics. For example, in Example 1.10, where we had a random sample from an exponential distribution, we found that $T = \sum X_i$ was sufficient for θ . However, (obviously) the whole data \underline{X} is sufficient for θ since

$$\begin{aligned} f_{\underline{X}}(\underline{x}|\theta) &= 1 \times f_{\underline{X}}(\underline{x}|\theta) \\ &= h(\underline{x}) g(\underline{x}, \theta), \end{aligned}$$

where $h(\underline{x}) = 1$ and $g(t, \theta) = f_{\underline{X}}(t|\theta)$. Also, any bijective (1-1) function of T is also sufficient for θ . For example, since $\bar{x} > 0$

$$\begin{aligned} f_{\underline{X}}(\underline{x}|\theta) &= 1 \times \theta^n \exp(-n\theta\bar{x}) \\ &= h(\underline{x}) g_1(\bar{x}, \theta), \text{ or} \\ &= h(\underline{x}) g_2((\bar{x})^2, \theta), \text{ or} \\ &= h(\underline{x}) g_3((\bar{x})^3, \theta), \text{ or} \\ &= h(\underline{x}) g_4((\bar{x})^4, \theta), \text{ or} \\ &= \dots \end{aligned}$$

where $h(\underline{x}) = 1$ and $g_j(t, \theta) = \theta^n \exp(-\theta t^{1/j})$, $j = 1, 2, \dots$. Therefore, $\sum X_i$ is sufficient for θ , but so are any of \underline{X} , \bar{X} , $(\bar{X})^2$, $(\bar{X})^3$, $(\bar{X})^4$, \dots , $\log(\bar{X})$, $\exp(\bar{X})$, \dots .

Further, if \underline{T} is sufficient for θ and \underline{S} is any other statistic then $(\underline{T}, \underline{S})$ is also sufficient for θ . For example,

$$\begin{aligned} f_{\underline{X}}(\underline{x}|\theta) &= 1 \times \theta^n \exp(-n\theta\bar{x}) \\ &= h(\underline{x}) g_1(\bar{x}, x_1, \theta), \text{ or} \\ &= h(\underline{x}) g_2(\bar{x}, x_2 \sin x_6, \theta), \text{ or} \\ &= \dots \end{aligned}$$

where $h(\underline{x}) = 1$ and $g_j(t_1, t_2, \theta) = \theta^n \exp(-\theta t_1)$.

Minimal Sufficiency

The main role of a sufficient statistic is to summarise information in the data about the parameters — this should be expressed as concisely as possible. Therefore, we want as few sufficient statistics as possible.

Definition 1.6

A statistic T is minimal sufficient if it is a function of every other sufficient statistic.

This definition does not uniquely define a minimally sufficient statistic, since any bijective function of a minimally sufficient statistic is also minimal sufficient. However, it does achieve the greatest reduction of the data without losing any information about the parameters.

Chapter 2

Bayes' Theorem for Distributions

2.1 Introduction

Suppose we have data \underline{x} which we model using the probability (density) function $f(\underline{x}|\theta)$. The likelihood function for θ is therefore $L(\theta|\underline{x}) = f(\underline{x}|\theta)$. 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.

Bayes' Theorem

The posterior probability (density) function for θ is

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

where

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

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

$$\pi(\theta|\underline{x}) \propto \pi(\theta) \times L(\theta|\underline{x})$$

i.e. **posterior \propto prior \times likelihood.**

Thus, to obtain the posterior distribution, we need:

- (1) data, from which we can form the likelihood, and
- (2) a suitable distribution that represents our prior beliefs about the parameter θ .

You should now be comfortable with how to obtain the likelihood (see Section 1.3 and MAS2302!). But how do we specify a prior? Before we consider the task of *prior elicitation* – the process from which we obtain a suitable prior distribution for θ – it will be useful to re-familiarise ourselves with two continuous distributions: the beta distribution and the gamma distribution (I will assume that, by now, you are more than familiar with other distributions, such as the exponential, Normal, Poisson, binomial...).

Definition 2.1 (Beta distribution)

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

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

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

$$B(a, b) = \int_0^1 y^{a-1}(1-y)^{b-1} dy. \quad (2.1)$$

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

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x} dx. \quad (2.2)$$

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

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

Also

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

Definition 2.2 (Gamma distribution)

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

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

where $\Gamma(a)$ is the gamma function defined in (2.2). It can be shown that

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

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 $Beta(2, 5)$, 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 $Beta(2, 5)$ 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 $Beta(6, 50)$ (top right), $Beta(71, 6)$ (bottom left) and $Beta(10, 10)$ (bottom right) distributions. Notice that different combinations of (a, b) give rise to different shapes of distribution, from positively skewed to 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`, `dpois` or `dnorm` instead of `dbeta` for the gamma, Poisson and Normal distributions, respectively.

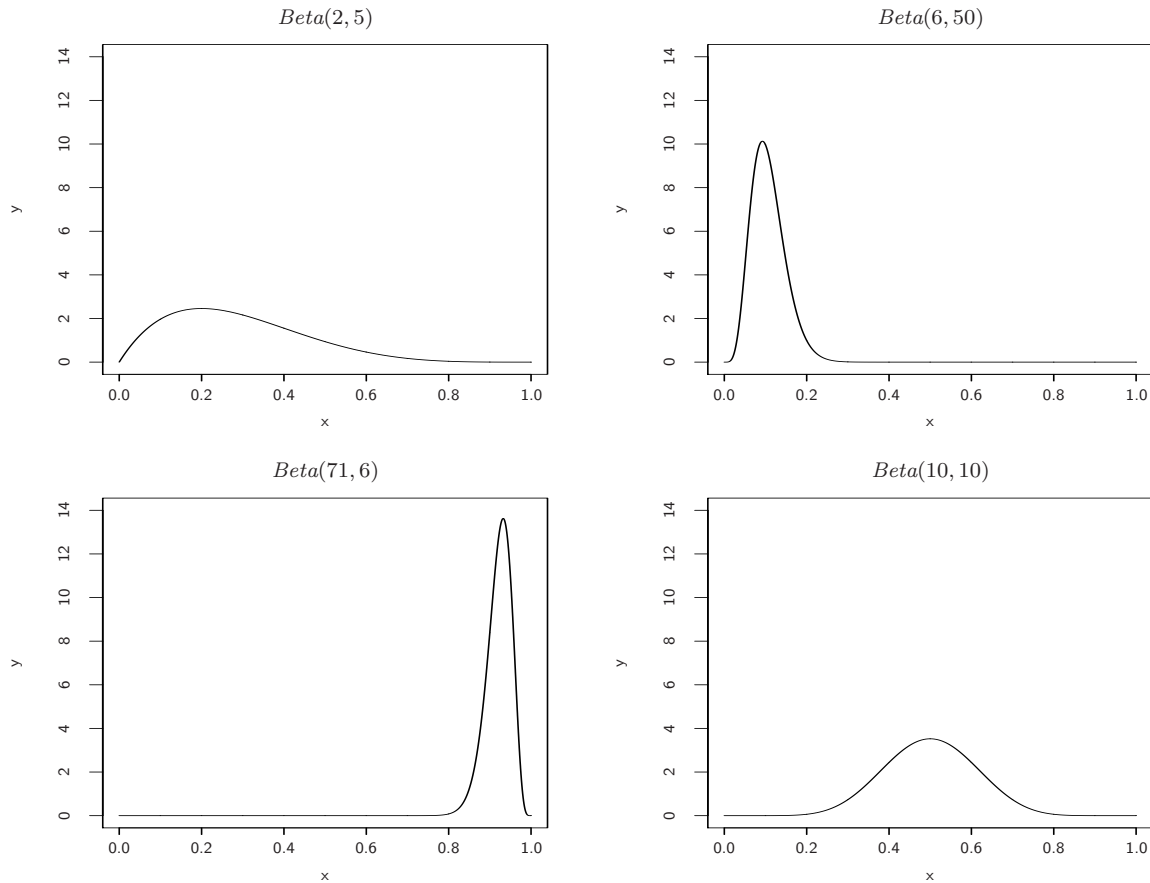



Figure 2.1: Plots of Beta densities for various values of (a, b) .

2.2 Prior elicitation

In this section we will think about how to construct suitable prior distributions $\pi(\theta)$ for various parameters of interest. Note that this is a huge area of recent – and current – research in Bayesian Statistics; the aim in this course is to give a brief (and simple!) overview.


Example 2.1

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* over the summer. Max has no idea about the level of interest in this game, and so can only assume that all values of θ are equally likely. Specify a suitable prior distribution for θ .

 ...Solution to Example 2.1...


Example 2.2

A *National Express* coach is due to arrive in Newcastle from London at 23:00. However, in practice, it is equally likely to arrive anywhere between 15 minutes early to 45 minutes late, depending on traffic conditions. Let θ denote the amount of time (in minutes) that the coach is delayed. Specify a suitable prior distribution for θ .


 ...Solution to Example 2.2...

Example 2.3

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. Specify a suitable prior for θ .

 *...Solution to Example 2.3...***Example 2.4**

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. In the space below, sketch what you think might be a suitable form for the prior distribution for θ . Why is a $\text{Uniform}(0, 1)$ not appropriate? Can you suggest a more appropriate statistical model for θ ?

 *...Solution to Example 2.4...*

In example 2.4 we are given some information about how likely certain values of θ are, and it is not too difficult to sketch what the prior distribution should look like. However, it is far more difficult a task to translate this into a useable statistical model. You should understand why a Uniform distribution is not appropriate here – we cannot assume that all values of θ are equally likely. You should also understand that some sort of $Beta(a, b)$ distribution might be a good candidate here. But how can we obtain suitable values of the parameters a and b ?

2.2.1 Elicitation using suggested prior summaries

If some simple prior summaries for θ can be specified – for example, its mean, mode or standard deviation – then it may be possible to use this information to elicit suitable parameters for our prior distribution $\pi(\theta)$.

Returning to example 2.4: we are told that the mode for θ should be around 0.95 and that $\Pr(\theta < 0.8)$ should be very small. Suppose further that we are told the mean should be around 0.92. Using the formulae on page 22, we then have

$$Mode = \frac{a - 1}{a + b - 2} = 0.95 \quad (2.3)$$

and

$$Mean = \frac{a}{a + b} = 0.92. \quad (2.4)$$

We can then solve these two expressions simultaneously to obtain (potentially) suitable values for a and b . Rearranging Equation (2.3), we have:

$$\begin{aligned} a - 1 &= 0.95a + 0.95b - 1.9 \\ \Rightarrow 0.05a &= 0.95b - 0.9 \\ \Rightarrow a &= 19b - 18. \end{aligned} \quad (2.5)$$

Substituting this into (2.4), we get

$$\begin{aligned} \frac{19b - 18}{19b - 18 + b} &= 0.92 \\ \Rightarrow 19b - 18 &= 18.4b - 16.56 \\ \Rightarrow 0.6b &= 1.44 \\ \Rightarrow b &= 2.4; \end{aligned}$$

substituting $b = 2.4$ into (2.5), we get

$$a = 19 \times 2.4 - 18 = 27.6,$$

suggesting a $Beta(27.6, 2.4)$ prior distribution for θ . R has been used as before to produce a plot of this distribution (Figure 2.2). Notice that this plot seems reasonable: it reflects the fact that the music expert is very competent, with most of the distribution being to the far right-hand side of the range of allowable values for θ . However, also notice the shaded area to the left of $\theta = 0.8$: we were told that $\Pr(\theta < 0.8)$ should be very small, and it could be that the shaded area in Figure 2.2 is too large. In fact, we can use R to work this out:

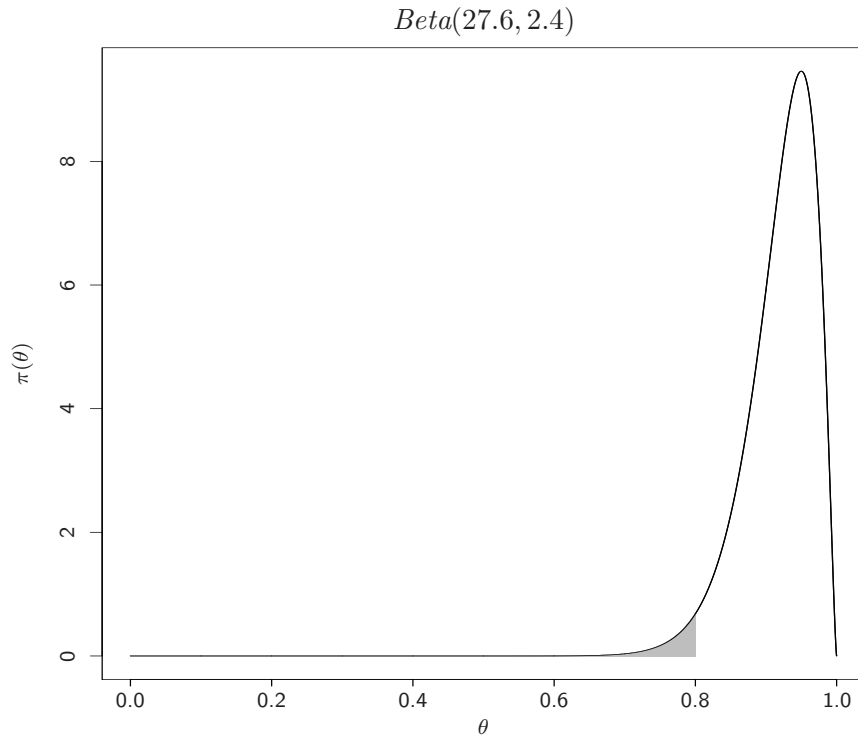


Figure 2.2: $Beta(27.6, 2.4)$ distribution, elicited from the specified mean and mode of 0.92 and 0.95, respectively.

```
> pbeta(0.8, 27.6, 2.4)
[1] 0.02381770
```

Perhaps we could refine our prior for θ such that the mean and mode match up with what they should (0.92 and 0.95, respectively), *and* $\Pr(\theta < 0.8)$ is even smaller than that under our current prior specification? Often, eliciting a prior distribution for a parameter θ involves several iterations of model development and feedback. For example, the statistician develops a potentially suitable prior for θ , as we have done above; this is then presented to the expert, who is then given the opportunity to comment on the suitability of our proposed model, perhaps checking things like quartiles or various probabilities, such as $\Pr(\theta < 0.8)$ in the above example. If s/he thinks these are not quite right, the statistician might then be given some further criteria which the prior should meet: for example, the expert might recommend that $\Pr(\theta < 0.8) = 0.0238$, as provided by our $Beta(27.6, 2.4)$ above, is be too high; perhaps s/he might recommend that this probability be no larger than 0.01 (0.1%), for example.

In the worked example given above, an expert has provided us with two measures of location for θ (the mean and mode), and some information about how likely it is that $\theta < 0.8$. However, it is often difficult for the expert to give us such precise information: for example, the expert might easily be able to tell what they think is the most likely value for θ , but should this be the prior mean or the prior mode? How might we, or the expert, be able to unravel such differences? Most experts might struggle to provide us with both a prior mean and a prior mode for θ .

Example 2.5

Let θ be the rate at which earthquakes occur, in days, along the boundary of the Eurasian plate (see Figure 2.3). A seismologist with particular expertise in this region tells us that we can expect earthquakes to occur at a rate of about 0.0025 per day (less than once per year, or about once every 400 days), with a variance of about 6.25×10^{-7} . Assuming $\theta \sim \text{Gamma}(a, b)$, find a and b and use R to visualise $\pi(\theta)$.


 ...Solution to Example 2.5...



Figure 2.3: Plates tectonics map by the *United States Geological Survey*.

2.2.2 Elicitation using the bisection method

Asking an expert to directly provide us with summaries of θ – for example the variance, or the mean and mode – can often be asking too much. A much more straightforward method of elicitation might be found via the *bisection method*. Suppose the parameter of interest – θ – is constrained to lie between 0 and 1 (for example, θ could be a probability/proportion). There are six main stages to the elicitation:

1. Elicit the expert's median, m

We ask the expert to choose a value m such that the two intervals $[0, m]$ and $[m, 1]$ have the same probability. In the music expert example, m would be much closer to 1 than to 0 because we are told that the expert is very competent.

What if the expert does not understand “have the same probability”? We could explain what is meant by this by describing a gamble in which, for a given m , the expert chooses one of the two intervals $[0, m]$ or $[m, 1]$, and receives a reward if θ lies in the chosen interval (but does not pay any penalty if θ lies in the other interval). If the expert judges the two intervals to have the same probability, then s/he would have no preference for one interval over the other in this gamble.

We can help the expert make this judgement by proposing values of m and asking the expert simply to consider which interval s/he judges to be more likely. For example, the expert may judge $[0, 0.5]$ to be more probable than $[0.5, 1]$, but judge $[0, 0.25]$ to be less probable than $[0.25, 1]$, implying that m must be somewhere in the interval $[0.25, 0.5]$.

2. Elicit the expert's lower quartile, l

We now ask the expert to divide the interval $[0, m]$ into two equally probable intervals: $[0, l]$ and $[l, m]$. In practice, the expert is likely to find this more difficult than choosing the median. We could help by asking the expert how certain s/he is about the value of θ – whether θ is very likely to be close to m , or whether it could be considerably lower. We could ask the expert if $[0, m/2]$ is more or less likely than $[m/2, m]$; we could then ask the expert to consider whether $[0, m - \delta]$ is more or less probable than $[m - \delta, m]$ (for some suitably small δ), leading the expert to consider a value for l in the interval $[m/2, m - \delta]$.

3. Elicit the expert's upper quartile, u

Next we elicit the expert's upper quartile u , noting the considerations in step 2, by asking him/her to divide the interval $[m, 1]$ into two equally probable intervals $[m, u]$ and $[u, 1]$.



4. Reflection

We should now invite the expert to reflect on his/her choices and check for consistency, by asking:

“Consider the four intervals: $[0, l]$, $[l, m]$, $[m, u]$ and $[u, 1]$. Do you consider any one of them to be more probable than any other?”

In theory, the expert’s answer should be “no”. If this is not the case, the expert should be asked to modify his/her choices of l , m and u .

5. Fit a parametric distribution to these judgements

We now fit a parametric distribution to the expert’s judgements. An obvious choice when θ lies between 0 and 1 is the *Beta*(a, b) distribution (Equation 2.1). We can obtain a and b using a *least squares approach* (i.e. see MAS2316); however, researchers at Sheffield University (Professor Tony O’Hagan and Dr. Jeremy Oakley) have developed a web-based program to help with this, called the *MATCH Uncertainty Elicitation Tool*. We will see this in action soon – it can be found at <https://optics.eee.nottingham.ac.uk/match/uncertainty.php>.

6. Feedback and refinement

Finally, we must check that the chosen distribution is an acceptable representation of the expert’s beliefs, given that the expert has only provided three quartiles (l , m and u). This is known as the ‘feedback stage’ and involves presenting the fitted distribution back to the expert, together with some additional summaries of the distribution. One option is to obtain the 0.33 and 0.66 quantiles of the proposed distribution which we denote by $\theta_{0.33}$ and $\theta_{0.66}$ (known as the *tertiles* of the distribution), so that $[0, \theta_{0.33}]$, $[\theta_{0.33}, \theta_{0.66}]$ and $[\theta_{0.66}, 1]$ are three equally probable intervals (the values of $\theta_{0.33}$ and $\theta_{0.66}$ can be obtained in *MATCH*). The expert may wish to specify alternative values for these tertiles, and/or modify his/her original judgements regarding the quartiles, in which we would need to return to step 5 (the ‘refinement stage’). Further feedback and refinement can be made until the expert is happy with the quantiles and tertiles.

As a final check, the statistician should discuss the tails of the proposed distribution with the expert, for example, by considering the 1st and 99th percentiles (again, we can use the *MATCH* software to obtain these values). Considering each tail in turn, the Statistician should ask the expert questions such as

*“What event would have to happen for θ to be this small/large?
How likely would such an event be?”*

If the expert is satisfied with the tails of the distribution, then the elicitation is concluded. Otherwise, the expert may wish to revise his/her initial judgements. The whole process is iterated until the expert is satisfied with the chosen fitted distribution.

Example 2.6

Over the past 15 years there has been considerable scientific interest in the rate of retreat, θ (metres per year), of glaciers in Greenland (as discussed in the recent *Frozen Planet* series shown on the BBC); indeed, this has often been used as an indicator of global warming. In this example we will use the bisection method, and the *MATCH* elicitation software, to suggest a suitable distribution for θ for the *Zachariae Isstrøm* glacier.

Records from an expert glaciologist show that glaciers in Greenland have been retreating at a rate of between 0 and 70 feet per year since 1995. We will use these values as the lower and upper limits for θ , respectively.

1. Elicit the expert's median, m

Statistician: “Can you give us a value m such that, for the *Zachariae Isstrøm* glacier, we can expect the rate of retreat in 2012 to have equal chance of lying in $[0, m]$ feet and $[m, 70]$ feet?”

Glaciologist: “Hmmm... the *Zachariae Isstrøm* glacier lies in quite a northerly region of glacial activity in Greenland, one that has not been *severely* affected by glacial retreat. For this glacier this year, there is a good chance that the rate of retreat will be lower than most in Greenland, perhaps around 24 feet.”

Statistician: “So for this glacier, it is just as likely that the rate of retreat will be somewhere in $[0, 24]$ and $[24, 70]$?”

Glaciologist: “Yes, I think that sounds reasonable!”

In *MATCH*, we enter the **Lower Limit** and **Upper Limit** as 0 and 70, respectively; after selecting the **Quartile** method of elicitation as the **Input Mode**, we then move the **Median** slider to 24: this can be seen in the screenshot in Figure 2.4. Notice that the green area of the median bar represents the lower 50% of the distribution for θ and the blue area represents the upper 50% of the distribution for θ . Notice also that the lower and upper quartiles are still at their default values – one quarter and three quarters of the range, respectively.

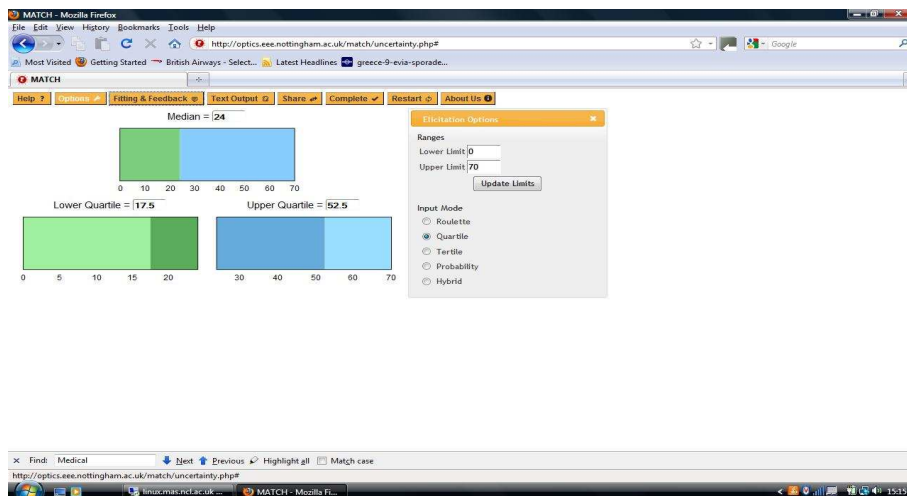


Figure 2.4:

2. Elicit the expert's lower quartile, l

Statistician: “So, you think there is an even chance that the *Zachariae Isstrøm* glacier will retreat between $[0, 24]$ feet and $[24, 70]$ feet this year. Can you split the lower interval $[0, 24]$ into two halves of equal probability also?”

Glaciologist: “Erm, not sure, that’s a bit more difficult...”

Statistician: “OK, you said you expect the glacier to retreat by about 24 feet this year. How certain are you of this value? Do you think it could be considerably lower than this?”

Glaciologist: “Well, obviously, I can’t be *certain*... but I doubt it would be *much* lower than this, for this particular glacier...”

Statistician: “Do you think $[0, 12]$ or $[12, 24]$ is more likely?”

Glaciologist: “Definitely, a rate of retreat somewhere between 12 feet and 24 feet is much more likely than between 0 and 12 feet. There are areas of Greenland further North where the glaciers have much slower rates of retreat... only the most northerly glaciers have zero retreat.”

Statistician: “OK. so $[12, 24]$ is more likely than $[0, 12]$. Is there a value between 12 and 24 that you’d be prepared to go down to for the rate of retreat for this glacier?”

Glaciologist: “Probably a bit more than half-way. Maybe 19 feet?”

3. Elicit the expert's upper quartile, u

Statistician: “Thank you. In a similar way, for this glacier, could you split the upper interval $[24, 70]$ into two halves of equal probability?” *Glaciologist:* “I’m more sure that the rate of retreat for this glacier will be closer to my specified value $[24]$ than half way between 24 and 70... really, only the fastest retreating glaciers in more southerly regions have a rate of retreat more than about 40 feet in one year... I think a value of about 30 feet would split this upper interval quite nicely here.”

Statistician: “Thank you.”

We now update the *MATCH* screen with the suggested lower and upper quartiles (19 and 30, respectively), before clicking **Fitting and Feedback**. Figure 2.5 shows a screenshot of this.

4. Reflection

Statistician: “So, would you consider the following four intervals equally likely?”

$$[0, 19], [19, 24], [24, 30], [30, 70]$$

Glaciologist: “This seems reasonable to me, I think...”

5. Fit a parametric distribution to these judgements

Notice that the screenshot from *MATCH*, shown in Figure 2.5, indicates that a $\text{Gamma}(9, 0.36)$ distribution might be appropriate for θ . Notice that *MATCH* gives the tertiles of this distribution as $\theta_{0.33} = 20.6$ and $\theta_{0.66} = 27.6$.

6. Feedback and refinement

We now show the glaciologist our distribution for the rate of glacial retreat, i.e. $\theta \sim \text{Gamma}(9, 0.36)$.

Statistician: “We have obtained a probability distribution for the rate of retreat for the *Zachariae Isstrøm* glacier. A plot of this distribution is shown in Figure 2.5. Further, this gives the following three intervals as being equally likely for the rate of retreat:

$$[0, 20.6], [20.6, 27.6], [27.6, 70]$$

Does this seem reasonable?”

Glaciologist: “So they all have probability one third?”

Statistician: “Yes, that’s right.”

Glaciologist: “I think this looks OK”

The Statistician should now consider the tails: we can do this by setting the **Feedback Percentiles** in *MATCH* to 1% and 99% (see Figure 2.6), giving $\theta_{0.01} = 9.74$ and $\theta_{0.99} = 48.2$.

Statistician: “What would have to happen for the rate of retreat at this glacier to be as much as 48 feet this year?”

Glaciologist: “Something pretty spectacular for a glacier at this longitude! Although it’s not *impossible*, just really, really unlikely.”

Statistician: “Our probability distribution gives this event, or anything more extreme, a probability of 0.01 – i.e. once in a hundred years – does this seem small enough?”

Glaciologist: “Yes, I suppose this is imaginable”

Statistician: “At the other end of the scale, we give a rate of retreat of just 10 feet this year the same small probability. Are you happy with this?”

Glaciologist: “Yes, that looks fine to me!”

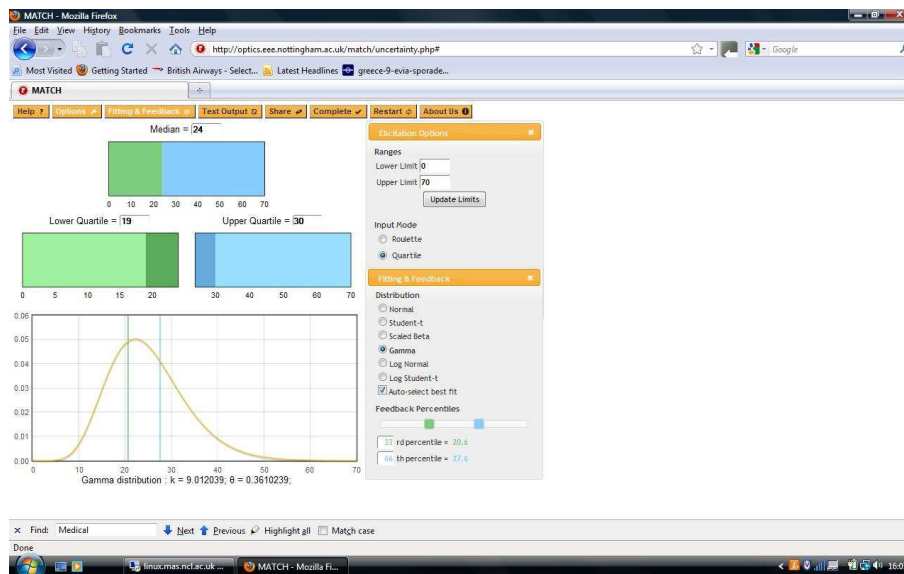


Figure 2.5:

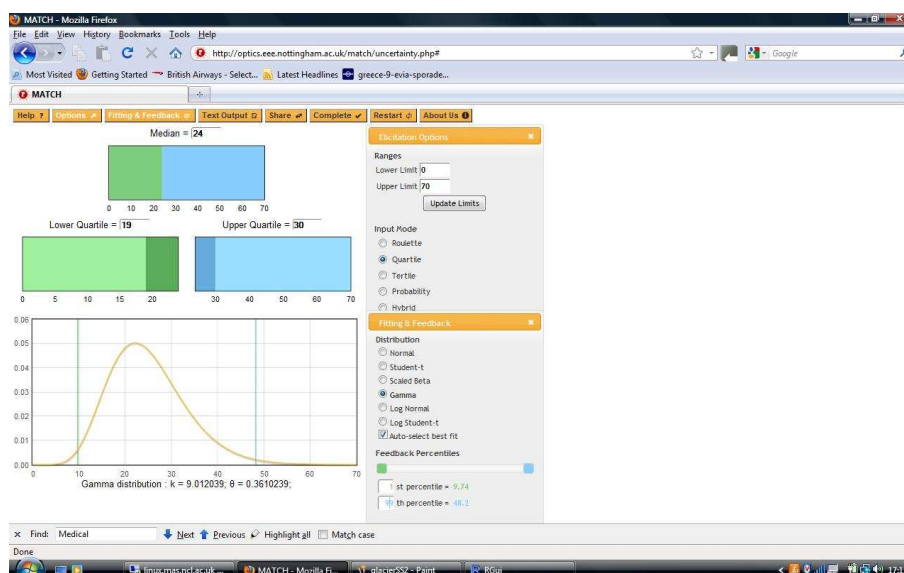


Figure 2.6:

Example 2.7

Let us return to the example of the music expert (Example 2.4). Recall that $\theta = \Pr(\text{correct choice})$ and we are given that $\text{Mode}(\theta) \approx 0.95$ and $\Pr(\theta < 0.8)$ should be very small. In Section 2.2.1 we were also told that $\text{Mean}(\theta) \approx 0.92$.

Let us try to elicit a suitable prior for θ using the bisection method. As in example 2.6, we use the *MATCH* software. Using steps 1–3 of the bisection method, and after careful reflection on the part of the expert (step 4), suppose we find that $l = 0.904$, $m = 0.926$ and $u = 0.944$ – i.e. there is a probability of 0.75 that θ lies in the interval $(0.904, 0.944)$. Figure 2.7 is a screenshot from *MATCH* showing the elicitation session.

MATCH suggests that $\theta \sim \text{Beta}(71, 6)$. The screenshot in Figure 2.7 also gives us the feedback percentiles – the defaults are set at the 33rd and 66th percentiles (i.e. the tertiles), and we could report these back to the expert, along with a picture of the elicited distribution, for checking and possible refinement.

We can visualise this distribution in R using commands like those on page 22 of these lecture notes. Doing so gives the plot in Figure 2.8. Notice that the shaded area to the left of 0.8 is smaller than that in Figure 2.2. In fact, we can work this out in R:

```
> pbeta(0.8, 71, 6)
[1] 0.001050360
```

Thus, we now have $\Pr(\theta < 0.8) = 0.0011$ compared to 0.0238 in the earlier elicitation using prior summaries only (Section 2.2.1). Since we are told that we want this probability to be very small, the current elicited beta distribution seems preferable to the earlier one; indeed, we could ask the expert what he/she thinks about this. Using the formulae on page 22, we also have

$$\text{Mode}(\theta) = \frac{71 - 1}{71 + 6 - 2} = 0.933 \quad \text{and}$$

$$E(\theta) = \frac{71}{71 + 6} = 0.922,$$

which match up with the required values (0.95 and 0.92) fairly closely.

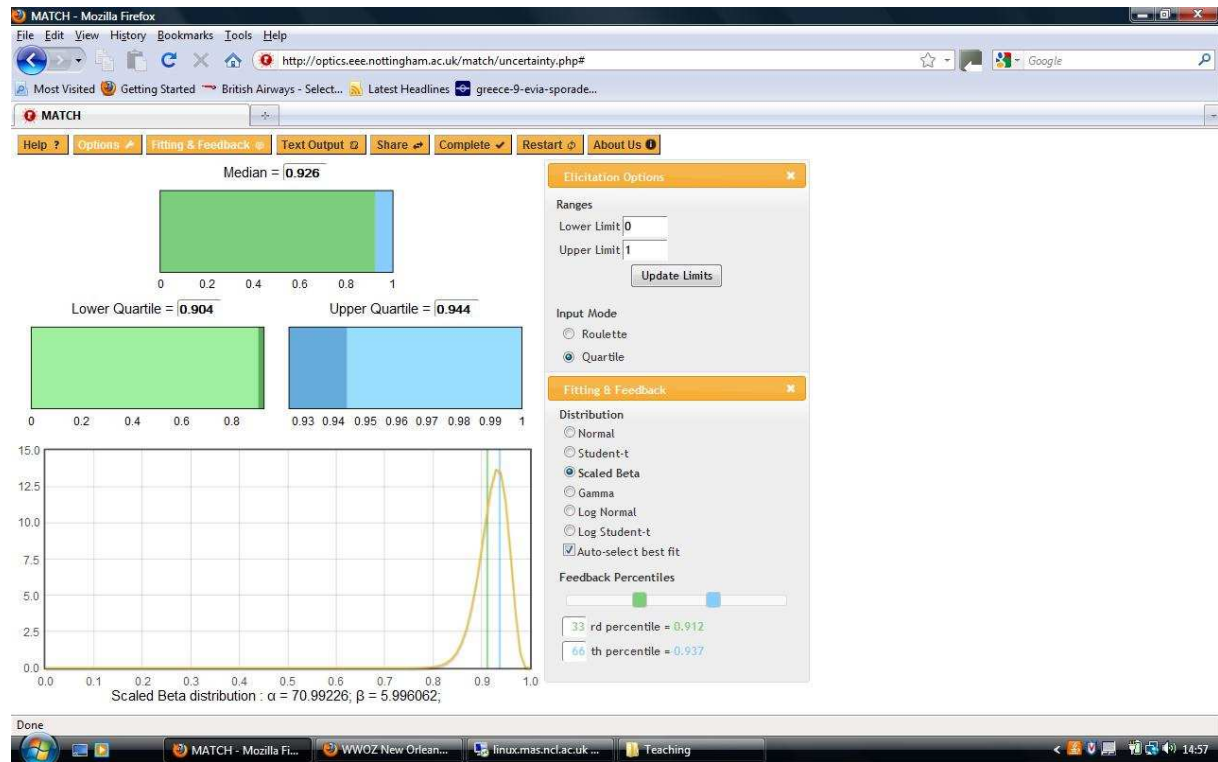


Figure 2.7: Screenshot from the *MATCH Uncertainty Elicitation Tool* for the music expert example: we see that a $Beta(71, 6)$ distribution has been suggested for θ .

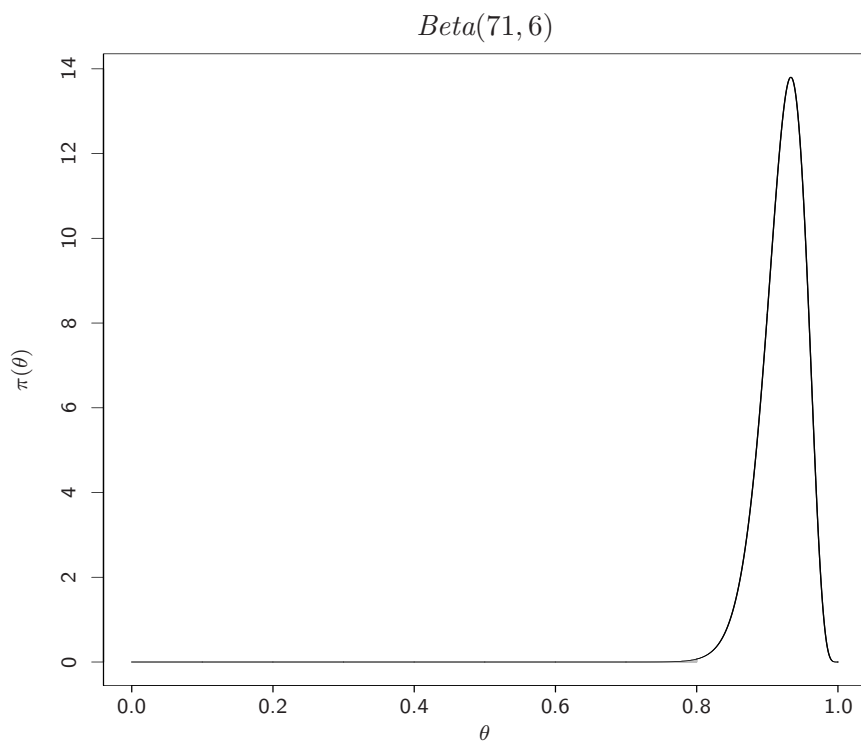


Figure 2.8: $Beta(71, 6)$ distribution, elicited using the bisection method in *MATCH*.

Example 2.8

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, giving a prior distribution $\theta \sim U(0, 1)$:

$$\pi(\theta) = 1, \quad 0 < \theta < 1. \quad (2.6)$$

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

The data is an observation on the random variable $X|\theta \sim \text{Bin}(5, \theta)$. This gives a likelihood function

$$L(\theta|x = 1) = f(x = 1|\theta) = 5\theta(1 - \theta)^4 \quad (2.7)$$

which favours values of θ near its maximum $\theta = 0.2$. Therefore, we have a conflict of opinions: the prior distribution (2.6) suggests that θ is probably around 0.5 and the data (2.7) suggest that it is around 0.2. We can use Bayes Theorem to combine these two sources of information in a coherent way. First

$$\begin{aligned} f(x = 1) &= \int_{\Theta} \pi(\theta)L(\theta|x = 1) d\theta & (2.8) \\ &= \int_0^1 1 \times 5\theta(1 - \theta)^4 d\theta \\ &= \int_0^1 \theta \times 5(1 - \theta)^4 d\theta \\ &= \left[-(1 - \theta)^5 \theta \right]_0^1 + \int_0^1 (1 - \theta)^5 d\theta \\ &= 0 + \left[-\frac{(1 - \theta)^6}{6} \right]_0^1 \\ &= \frac{1}{6}. \end{aligned}$$

Therefore, the posterior density is

$$\begin{aligned}
 \pi(\theta|x=1) &= \frac{\pi(\theta)L(\theta|x=1)}{f(x=1)} \\
 &= \frac{5\theta(1-\theta)^4}{1/6}, & 0 < \theta < 1 \\
 &= 30\theta(1-\theta)^4, & 0 < \theta < 1 \\
 &= \frac{\theta(1-\theta)^4}{B(2,5)}, & 0 < \theta < 1,
 \end{aligned}$$

and so the posterior distribution is $\theta|x=1 \sim \text{Beta}(2,5)$ – see Definition 2.1. This distribution has its mode at $\theta = 0.2$, and mean at $E[\theta|x=1] = 2/7 = 0.286$.

The main difficulty in calculating the posterior distribution was in obtaining the $f(x)$ term (2.8). 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

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

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 $\text{Beta}(g,h)$ distribution. Therefore, choosing g and h appropriately, we can see that the posterior distribution is $\theta|x=1 \sim \text{Beta}(2,5)$.

Summary:

It is possible that we have a biased coin. If we suppose that all values of $\theta = \text{Pr}(\text{Head})$ 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 2.9.

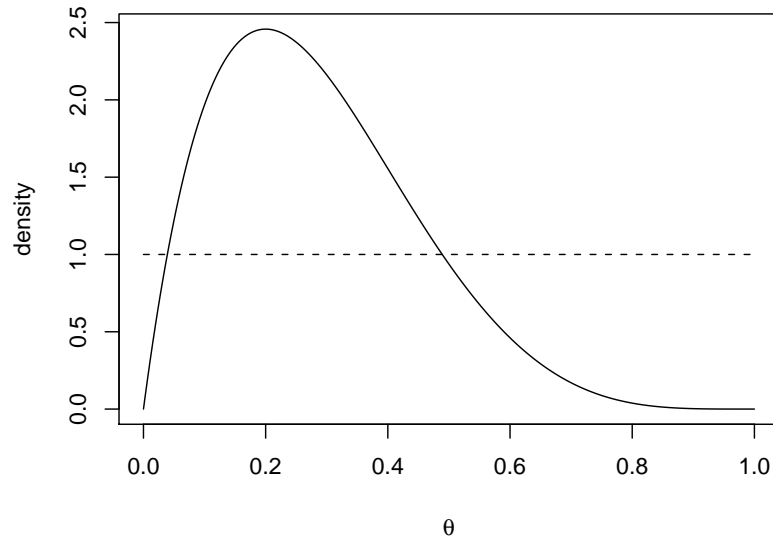


Figure 2.9: Prior (dashed) and posterior (solid) densities for $\theta = \text{Pr}(\text{Head})$

Example 2.9

Consider an experiment to determine how good a music expert is at distinguishing between pages from Haydn and Mozart scores. Let $\theta = \text{Pr}(\text{correct choice})$. 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 $\text{Pr}(\theta < 0.8)$ is very small. We choose $\theta \sim \text{Beta}(70, 6)$, with probability density function

$$\pi(\theta) = 128107980 \theta^{76} (1 - \theta)^4, \quad 0 < \theta < 1. \quad (2.9)$$

A graph of this prior density is given in Figure 2.10. 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

We have an observation on the random variable $X | \theta \sim \text{Bin}(10, \theta)$. This gives a likelihood function of

$$L(\theta | x = 9) = f(x = 9 | \theta) = 10 \theta^9 (1 - \theta) \quad (2.10)$$

which favours values of θ near its maximum $\theta = 0.9$. We combine these two sources of information using

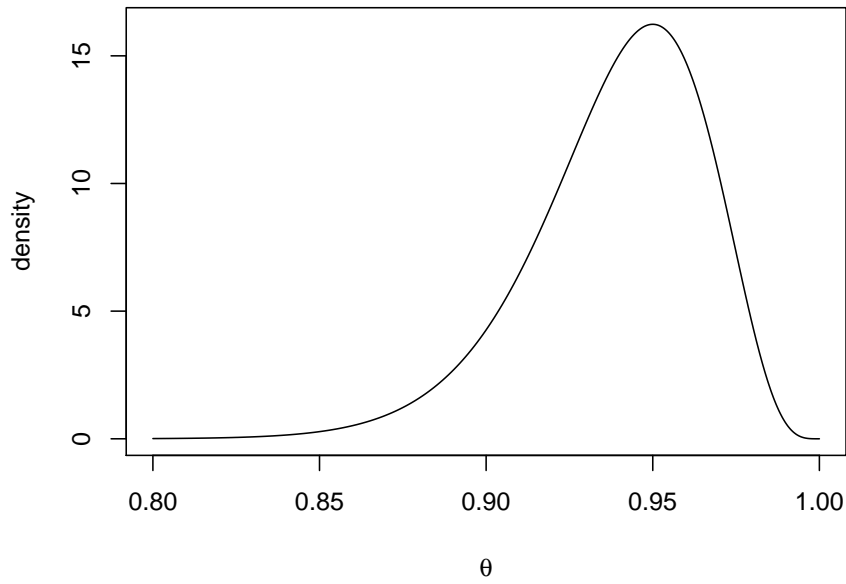


Figure 2.10: Prior density for the music expert's skill

Bayes Theorem. The posterior density function is

$$\begin{aligned}
 \pi(\theta|x=9) &\propto \pi(\theta)L(\theta|x=9) \\
 &\propto 128107980 \theta^{76}(1-\theta)^4 \times 10 \theta^9(1-\theta), \quad 0 < \theta < 1 \\
 &= k\theta^{85}(1-\theta)^5, \quad 0 < \theta < 1. \quad (2.11)
 \end{aligned}$$

We can recognise this density function as one from the Beta family. Whence, the posterior distribution is $\theta|x=9 \sim \text{Beta}(86, 6)$.

Summary:

The changes in our beliefs about θ are described by the prior and posterior distributions shown in Figure 2.11 and summarised in Table 2.1.

	Prior (2.9)	Likelihood (2.10)	Posterior (2.11)
$Mode(\theta)$	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

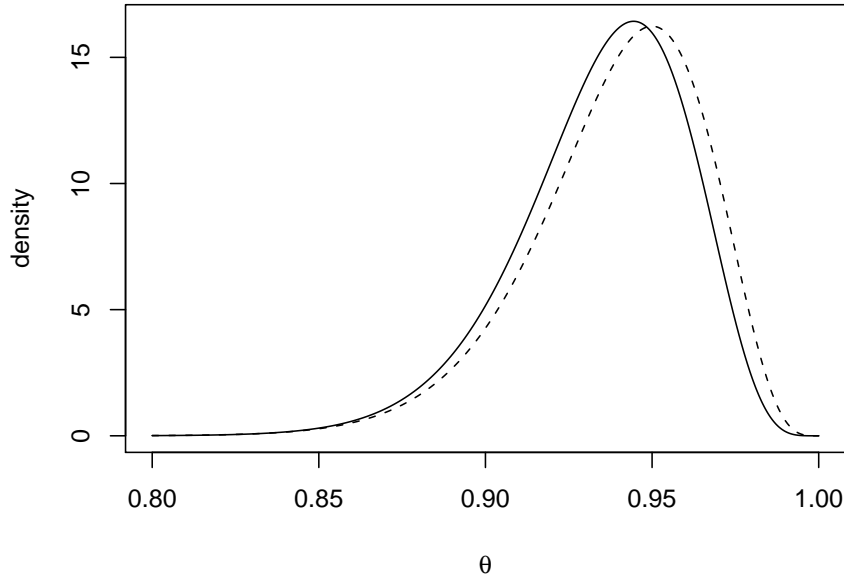


Figure 2.11: Prior (dashed) and posterior (solid) densities for the music expert's skill

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.

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

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

where $\Gamma(a)$ is the gamma function defined in (2.2). It can be shown that

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

Also

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

Example 2.10

Table 2.2 shows some data on the times between serious earthquakes world-wide. 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 Turkestan). The

table includes data on 21 earthquakes, that is, 20 “waiting times” between earthquakes. It is believed that earthquakes happen in a random haphazard kind of way and that

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

Table 2.2: Time intervals between major earthquakes (in days)

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, with density

$$\pi(\theta) = \frac{4000^{10} \theta^9 e^{-4000\theta}}{\Gamma(10)}, \quad \theta > 0, \quad (2.12)$$

and mean $E(\theta) = 0.0025$. A plot of this prior distribution can be found in Figure 2.12. 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 .

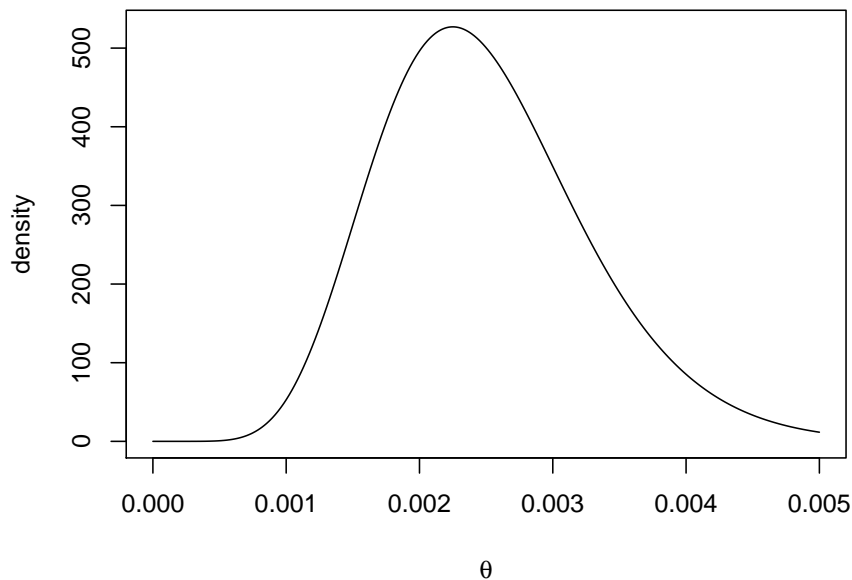


Figure 2.12: Prior density for the earthquake rate θ

Solution

The data are observations on $X_i|\theta \sim \text{Exp}(\theta)$, $i = 1, 2, \dots, 20$ (independent). Therefore, the likelihood function for θ is

$$\begin{aligned} L(\theta|\underline{x}) = f(\underline{x}|\theta) &= \prod_{i=1}^{20} \theta e^{-\theta x_i}, \quad \theta > 0 \\ &= \theta^{20} \exp \left(-\theta \sum_{i=1}^{20} x_i \right), \quad \theta > 0 \\ &= \theta^{20} e^{-9633\theta}, \quad \theta > 0. \end{aligned} \quad (2.13)$$

We now apply Bayes Theorem to combine the expert opinion with the observed data. The posterior density function is

$$\begin{aligned} \pi(\theta|\underline{x}) &\propto \pi(\theta) L(\theta|\underline{x}) \\ &\propto \frac{4000^{10} \theta^9 e^{-4000\theta}}{\Gamma(10)} \times \theta^{20} e^{-9633\theta}, \quad \theta > 0 \\ &= k \theta^{30-1} e^{-13633\theta}, \quad \theta > 0. \end{aligned} \quad (2.14)$$

The only continuous distribution which takes the form $k\theta^{g-1}e^{-h\theta}$, $\theta > 0$ is the $Ga(g, h)$ distribution. Therefore, the posterior distribution must be $\theta|\underline{x} \sim Ga(30, 13633)$.

Thus 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 2.13, and Table 2.3 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 .

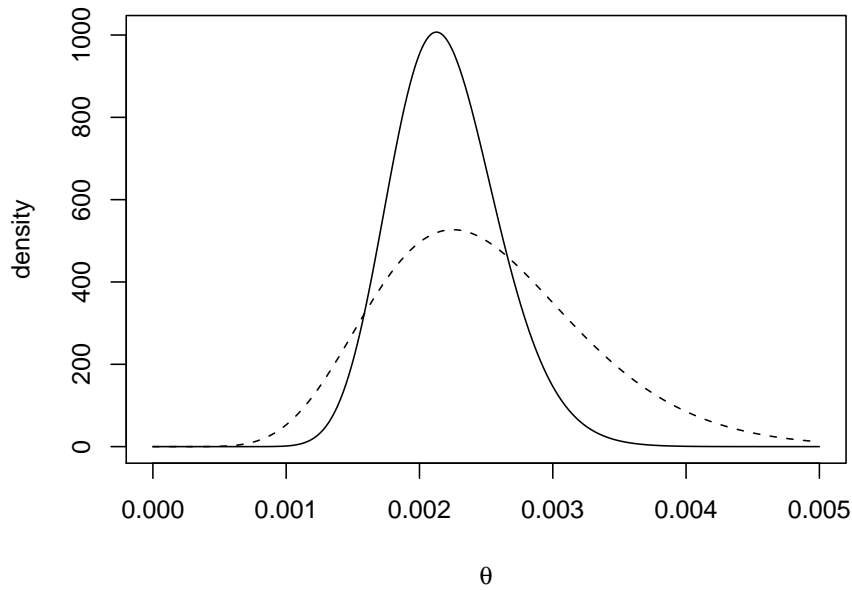


Figure 2.13: Prior (dashed) and posterior (solid) densities for the earthquake rate θ

	Prior (2.12)	Likelihood (2.13)	Posterior (2.14)
$Mode(\theta)$	0.00225	0.00208	0.00213
$E(\theta)$	0.00250	—	0.00220
$SD(\theta)$	0.00079	—	0.00040

Table 2.3: Changes in beliefs about θ

Example 2.11

We now consider the general case of the problem discussed in Example 2.10. Suppose $X_i|\theta \sim \text{Exp}(\theta)$, $i = 1, 2, \dots, n$ (independent) and our prior beliefs about θ are summarised by a $Ga(g, h)$ distribution (with g and h known), with density

$$\pi(\theta) = \frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \quad (2.15)$$

Determine the posterior distribution for θ .

Solution

The likelihood function for θ is

$$\begin{aligned} L(\theta|\underline{x}) = f(\underline{x}|\theta) &= \prod_{i=1}^n \theta e^{-\theta x_i}, \quad \theta > 0 \\ &= \theta^n e^{-n\bar{x}\theta}, \quad \theta > 0. \end{aligned} \quad (2.16)$$

We now apply Bayes Theorem. The posterior density function is

$$\begin{aligned} \pi(\theta|\underline{x}) &\propto \pi(\theta)L(\theta|\underline{x}) \\ &\propto \frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \theta^n e^{-n\bar{x}\theta}, \quad \theta > 0. \\ \pi(\theta|\underline{x}) &= k\theta^{g+n-1} e^{-(h+n\bar{x})\theta}, \quad \theta > 0. \end{aligned} \quad (2.17)$$

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

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 2.4, 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,

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

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

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

	Prior (2.15)	Likelihood (2.16)	Posterior (2.17)
$Mode(\theta)$	$(g-1)/h$	$1/\bar{x}$	$(g+n-1)/(h+n\bar{x})$
$E(\theta)$	g/h	–	$(g+n)/(h+n\bar{x})$
$SD(\theta)$	\sqrt{g}/h	–	$\sqrt{g+n}/(h+n\bar{x})$

Table 2.4: Changes in beliefs about θ **Example 2.12**

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 ($= 1/\text{variance}$) 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

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}. \quad (2.18)$$

Determine the posterior distribution for μ .

Solution

The likelihood function for μ is

$$\begin{aligned}
L(\mu|\underline{x}) &= f(\underline{x}|\mu) = \prod_{i=1}^n \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x_i - \mu)^2\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2} \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right]\right\}
\end{aligned}$$

Let

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

and then

$$L(\mu|\underline{x}) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp \left\{ -\frac{n\tau}{2} [s^2 + (\bar{x} - \mu)^2] \right\}. \quad (2.19)$$

Applying Bayes Theorem, the posterior density function is

$$\begin{aligned} \pi(\mu|\underline{x}) &\propto \pi(\mu)L(\mu|\underline{x}) \\ &\propto \left(\frac{d}{2\pi}\right)^{1/2} \exp \left\{ -\frac{d}{2}(\mu - b)^2 \right\} \\ &\quad \times \left(\frac{\tau}{2\pi}\right)^{n/2} \exp \left\{ -\frac{n\tau}{2} [s^2 + (\bar{x} - \mu)^2] \right\} \\ &= k_1 \exp \left\{ -\frac{1}{2} [d(\mu - b)^2 + n\tau(\bar{x} - \mu)^2] \right\} \end{aligned}$$

where k_1 is a constant that does not depend on μ . Now the exponent can be simplified by expanding terms in μ and then completing the square, as follows.

We have

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

where c does not depend on μ . Let

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \quad \text{and} \quad D = d + n\tau. \quad (2.20)$$

Then

$$\begin{aligned} \pi(\mu|\underline{x}) &= k_1 \exp \left\{ -\frac{D}{2}(\mu - B)^2 - \frac{c}{2} \right\} \\ &= k \exp \left\{ -\frac{D}{2}(\mu - B)^2 \right\}, \end{aligned} \quad (2.21)$$

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

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.

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

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

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

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

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

The changes in our beliefs about μ are summarised in Table 2.5. 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

$$E(\mu|\underline{x}) > E(\mu) \quad \Longleftrightarrow \quad \text{Mode}[L(\mu|\underline{x})] > E(\mu).$$

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

	Prior (2.18)	Likelihood (2.19)	Posterior (2.21)
$Mode(\mu)$	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 2.5: Changes in beliefs about μ **Example 2.13**

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

We have $n = 1$, $\bar{x} = x = 421$, $\tau = 1/64$, $b = 370$ and $d = 1/400$. Therefore, using the results in Example 2.12

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} = \frac{370/400 + 421/64}{1/400 + 1/64} = 414.0$$

$$D = d + n\tau = 1/400 + 1/64 = 1/7.43^2$$

and so the posterior distribution is $\mu|x = 421 \sim N(414.0, 7.43^2)$.

The (posterior) probability that the rock will be older than 400 million years is

$$Pr(\mu > 400|x = 421) = 0.9702$$

– calculated using the R commands

`1-pnorm(400,414,7.43)` or `1-pnorm(-1.884)`
or `pnorm(1.884)`.

Without the chemical analysis, the only basis for determining the age of the rock is via the prior distribution: the (prior) probability that the rock will be older than 400 million years is

$$Pr(\mu > 400) = 0.0668$$

— calculated using the R command `1-pnorm(400,370,20)`.

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 2.14.

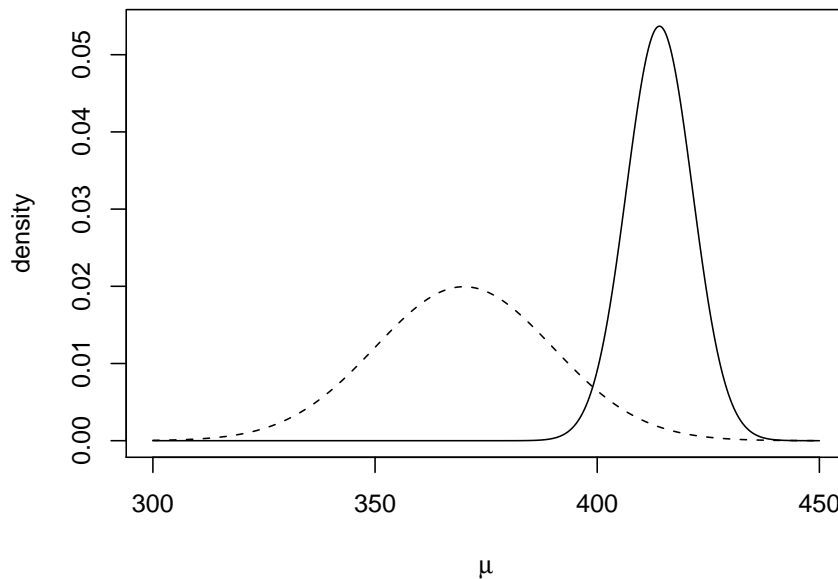


Figure 2.14: Prior (dashed) and posterior (solid) densities for the age of the rock

2.3 Sufficiency

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 $\underline{X} = (X_1, X_2, \dots, X_n)^T$ available and we want to make inferences about the parameters θ in the statistical model $f_{\underline{X}}(\underline{x}|\theta)$. If \underline{T} is a set of minimal sufficient statistics then by the Factorisation Theorem

$$\begin{aligned} L(\theta|\underline{x}) &= f_{\underline{X}}(\underline{x}|\theta) \\ &= h(\underline{x}) g(\underline{T}, \theta) \text{ for some functions } h \text{ and } g. \end{aligned}$$

Therefore, using Bayes Theorem

$$\begin{aligned} \pi(\theta|\underline{x}) &\propto \pi(\theta) L(\theta|\underline{x}) \\ &\propto \pi(\theta) h(\underline{x}) g(\underline{T}, \theta) \\ &\propto \pi(\theta) g(\underline{T}, \theta). \end{aligned}$$

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

$$g(\underline{T}, \theta) \propto f_{\underline{T}}(\underline{t}|\theta).$$

Hence

$$\pi(\theta|\underline{x}) \propto \pi(\theta) f_{\underline{T}}(\underline{t}|\theta).$$

However, applying Bayes Theorem to the data \underline{t} gives

$$\pi(\theta|\underline{t}) \propto \pi(\theta) f_{\underline{T}}(\underline{t}|\theta)$$

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

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

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

Example 2.14

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

Solution

The density of the data is

$$\begin{aligned}
 f_{\underline{X}}(\underline{x}|\theta) &= \prod_{i=1}^n \theta e^{-\theta x_i} \\
 &= \theta^n \exp \left(-\theta \sum_{i=1}^n x_i \right) \\
 &= 1 \times \theta^n \exp \left(-\theta \sum_{i=1}^n x_i \right) \\
 &= h(\underline{x}) g(\sum x_i, \theta)
 \end{aligned}$$

and therefore, by the Factorisation Theorem, $T = \sum_{i=1}^n X_i$ is sufficient for θ . Now $T|\theta \sim Ga(n, \theta)$ and so

$$L(\theta|t) = f_T(t|\theta) = \frac{\theta^n t^{n-1} e^{-\theta t}}{\Gamma(n)}, \quad \theta > 0.$$

Also

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

Therefore, by Bayes Theorem

$$\begin{aligned}
 \pi(\theta|t) &\propto \pi(\theta) L(\theta|t) \\
 &\propto \frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \frac{\theta^n t^{n-1} e^{-\theta t}}{\Gamma(n)}, \quad \theta > 0 \\
 &\propto \theta^{g+n-1} e^{-(h+t)\theta}, \quad \theta > 0
 \end{aligned}$$

and so the posterior distribution is $\theta|t \sim Ga(g+n, h+t)$. This is the same as the result we obtained previously for $\theta|\underline{x}$.

Example 2.15

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

Solution

Recall from (2.19) that

$$\begin{aligned} f_{\underline{X}}(\underline{x}|\mu) &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp \left\{ -\frac{n\tau}{2} [s^2 + (\bar{x} - \mu)^2] \right\} \\ &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp \left\{ -\frac{n\tau s^2}{2} \right\} \times \exp \left\{ -\frac{n\tau}{2} (\bar{x} - \mu)^2 \right\} \\ &= h(\underline{x}) g(\bar{x}, \mu) \end{aligned}$$

and therefore, by the Factorisation Theorem, $T = \bar{X}$ is sufficient for μ . Now $T|\mu \sim N(\mu, 1/(n\tau))$ and so

$$L(\mu|t) = f_T(t|\mu) = \left(\frac{n\tau}{2\pi}\right)^{1/2} \exp \left\{ -\frac{n\tau}{2} (t - \mu)^2 \right\}.$$

Also

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

Therefore, by Bayes Theorem

$$\begin{aligned}
\pi(\mu|t) &\propto \pi(\mu)L(\mu|t) \\
&\propto \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \\
&\quad \times \left(\frac{n\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{n\tau}{2}(t - \mu)^2\right\} \\
&\propto \exp\left\{-\frac{d}{2}(\mu - b)^2 - \frac{n\tau}{2}(t - \mu)^2\right\} \\
&\quad \vdots \\
&\propto \exp\left\{-\frac{D}{2}(\mu - B)^2\right\}
\end{aligned}$$

where B and D are as in (2.20), with t replacing \bar{x} ; that is, $\mu|t \sim N(B, 1/D)$, the same distribution as $\mu|\underline{x}$.

Definition 2.4

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

$$f(y|a, b, c) = \frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{ac}\Gamma\left(\frac{a}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left\{1 + \frac{(y-b)^2}{ac}\right\}^{-\frac{a+1}{2}}, \quad -\infty < y < \infty,$$

where $\Gamma(a)$ is the gamma function defined in (2.2). This distribution is a generalisation of Student's t -distribution since $(Y - b)/\sqrt{c} \sim t_a$. It can be shown that

$$E(Y) = \text{Mode}(Y) = b \quad \text{and} \quad \text{Var}(Y) = \frac{ac}{a-2}, \quad \text{if } a \geq 2.$$

Example 2.16

Suppose we have a random sample from a normal distribution in which both the mean μ and the precision τ are unknown, that is, $X_i|\mu, \tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \dots, n$ (independent). We shall adopt a prior distribution for (μ, τ) for which

$$\mu|\tau \sim N\left(b, \frac{1}{c\tau}\right) \quad \text{and} \quad \tau \sim Ga(g, h)$$

for known values b, c, g and h . We write $(\mu, \tau) \sim NGa(b, c, g, h)$ and this distribution has density function

$$\begin{aligned}\pi(\mu, \tau) &= \pi(\mu|\tau)\pi(\tau) \\ &= \left(\frac{c\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{c\tau}{2}(\mu - b)^2\right\} \times \frac{h^g \tau^{g-1} e^{-h\tau}}{\Gamma(g)}, \quad -\infty < \mu < \infty, \tau > 0 \\ &\propto \tau^{g-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}[c(\mu - b)^2 + 2h]\right\}, \quad -\infty < \mu < \infty, \tau > 0.\end{aligned}\quad (2.22)$$

Determine the posterior distribution for (μ, τ) .

Solution

Previously we have seen that the likelihood function is

$$L(\mu, \tau|\underline{x}) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{n\tau}{2}\{s^2 + (\bar{x} - \mu)^2\}\right].$$

Using Bayes Theorem, the posterior density is

$$\pi(\mu, \tau|\underline{x}) \propto \pi(\mu, \tau)L(\mu, \tau|\underline{x})$$

and so, for $-\infty < \mu < \infty, \tau > 0$

$$\begin{aligned}\pi(\mu, \tau|\underline{x}) &\propto \tau^{g-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}[c(\mu - b)^2 + 2h]\right\} \\ &\quad \times \tau^{\frac{n}{2}} \exp\left[-\frac{n\tau}{2}\{s^2 + (\bar{x} - \mu)^2\}\right] \\ &\propto \tau^{g+\frac{n-1}{2}} \exp\left\{-\frac{\tau}{2}[c(\mu - b)^2 + 2h + ns^2 + n(\bar{x} - \mu)^2]\right\} \\ &\quad \vdots \\ &\propto \tau^{G-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}[C(\mu - B)^2 + 2H]\right\}\end{aligned}\quad (2.23)$$

where

$$\begin{aligned} B &= \frac{bc + n\bar{x}}{c + n}, & C &= c + n, \\ G &= g + \frac{n}{2}, & H &= h + \frac{cn(\bar{x} - b)^2}{2(c + n)} + \frac{ns^2}{2}. \end{aligned} \quad (2.24)$$

Notice that the posterior distribution (2.23) is of the same form as the prior distribution (2.22). Therefore, we can conclude that the posterior distribution is

$$(\mu, \tau) | \underline{x} \sim NGa(B, C, G, H),$$

that is

$$\mu | \tau, \underline{x} \sim N \left(B, \frac{1}{C\tau} \right) \quad \text{and} \quad \tau | \underline{x} \sim Ga(G, H).$$

Comment

In this example, both the prior and posterior distributions are NGa distributions. We can determine the (marginal) means and variances of this distribution as follows. Suppose $(\mu, \tau) \sim NGa(b, c, g, h)$. Then we know that $\tau \sim Ga(g, h)$ and so the mean and variance of τ are easily determined. However, to calculate the mean and variance of μ we must first determine the marginal distribution for μ :

$$\begin{aligned} \pi(\mu) &= \int_0^\infty \pi(\mu, \tau) d\tau \\ &\propto \int_0^\infty \tau^{g-\frac{1}{2}} \exp \left\{ -\frac{\tau}{2} [c(\mu - b)^2 + 2h] \right\} d\tau. \end{aligned}$$

Now, as the integral of a gamma density over its entire range is one, we have

$$\int_0^\infty \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)} d\theta = 1 \quad \implies \quad \int_0^\infty \theta^{a-1} e^{-b\theta} d\theta = \frac{\Gamma(a)}{b^a}.$$

Therefore

$$\begin{aligned}\pi(\mu) &\propto \int_0^\infty \tau^{g+\frac{1}{2}-1} \exp\left\{-\frac{\tau}{2} [c(\mu-b)^2 + 2h]\right\} d\tau \\ &\propto \frac{\Gamma(g+\frac{1}{2})}{[\{c(\mu-b)^2 + 2h\}/2]^{g+\frac{1}{2}}} \\ &\propto \left\{1 + \frac{c(\mu-b)^2}{2h}\right\}^{-\frac{2g+1}{2}}.\end{aligned}$$

Hence

$$\mu \sim St\left(2g, b, \frac{h}{gc}\right). \quad (2.25)$$

Thus, marginally, μ has a generalised Student distribution and so we can determine the mean and variance of μ from known results about the mean and variance of this generalised Student distribution.

Returning to the example, notice how the information from the data combines with our prior beliefs to produce posterior beliefs about μ and τ .

The posterior mean of μ is greater than its prior mean if and only if the sample mean (likelihood mode) is greater than its prior mean, that is,

$$E(\mu|\underline{x}) > E(\mu) \quad \Longleftrightarrow \quad \bar{x} > b.$$

The relationships between the prior and posterior variance of μ and mean and variance of τ is rather more complex.

Example 2.17

Recall Example 2.13 on the ages of *Ennerdale granophyre* rocks. Previously we assumed that the ages followed a $N(\mu, 8^2)$ distribution, that is, the standard deviation of the age distribution was known to be 8. Now we consider the case where this standard deviation is unknown and determine posterior distributions using the theory in Example 2.16.

Before we can proceed, we must specify the parameters in the $NGa(b, c, g, h)$ prior distribution for (μ, τ) . In the previous analysis, we assumed that the population measurement precision was $\tau = 1/8^2 = 1/64$ and assumed a $N(370, 20^2)$ prior distribution for the population mean, that is, $\mu|\tau = 1/64 \sim N(370, 20^2)$.

Choice of b and c : the conditional prior distribution for μ is $\mu|\tau \sim N(b, 1/(c\tau))$ and so matching the prior distributions for μ (when $\tau = 1/64$) gives $b = 370$ and $c = 0.16$.

Choice of g and h : the marginal prior distribution for τ is $\tau \sim Ga(g, h)$. Previously, we assumed $\tau = 1/64$ (with $Var(\tau) = 0$) and so take this value as the prior mean: $E(\tau) = 1/64$. Also we will take $Var(\tau) = 1/128^2$. These two requirements give $g = 4$ and $h = 256$. Therefore, we will assume the prior distribution

$$(\mu, \tau) \sim NGa(370, 0.16, 4, 256).$$

We have seen that if $(\mu, \tau) \sim NGa(b, c, g, h)$ then the marginal distribution of μ is $\mu \sim St(2g, b, h/(gc))$. Therefore, with this choice of prior distribution, the marginal prior distribution for μ is

$$\mu \sim St(8, 370, 400).$$

Figure 2.15 shows the close match between the new (marginal) prior distribution for μ and that used previously. Figures 2.16 and 2.17 show the (marginal) distribution for τ and the corresponding distribution for $\sigma = 1/\sqrt{\tau}$.

We can combine the information in the $NGa(370, 0.16, 4, 256)$ prior distribution for (μ, τ) with that in the data ($n = 1$, $\bar{x} = x = 421$, $s^2 = 0$) using the results in Example 2.16 to obtain a $NGa(B, C, G, H)$ posterior distribution, where

$$\begin{aligned} B &= \frac{bc + n\bar{x}}{c + n} = \frac{(370 \times 0.16) + (1 \times 421)}{1.16} = 414.0, \\ C &= c + n = 1.16, \\ G &= g + \frac{n}{2} = 4.5, \\ H &= h + \frac{cn(\bar{x} - b)^2}{2(c + n)} + \frac{ns^2}{2} = 256 + \frac{0.16}{2.32}(421 - 370)^2 + 0 = 435.38. \end{aligned}$$

Plots of the (marginal) prior and posterior distributions of μ and τ are given in Figures 2.18 and 2.19. We can determine the (marginal) prior and posterior distributions for σ from that of τ ; see Figure 2.20. We can also examine the joint prior and posterior distributions for (μ, τ) via their contour plots to see if there is any change in the dependence structure; see Figure 2.21. The shape of the contours for the posterior distribution are fairly similar to those for the prior distribution and so the dependence structure has changed little. The main changes shown by the figure are in the mean and variability of μ and τ .

The (posterior) probability that the rock will be older than 400 million years is calculated as follows. Using (2.25), we have $\mu|\underline{x} \sim St(9, 414.0, 83.406)$ and so $(\mu - 414.0)/\sqrt{83.406}|\underline{x} \sim t_9$. Therefore

$$Pr(\mu > 400|x = 421) = 0.9202$$

– calculated using the R command `1-pt((400-414)/sqrt(83.406), 9)`.

Without the chemical analysis, the only basis for determining the age of the rock is via the prior distribution. Here the prior distribution is $\mu \sim St(8, 370, 400)$, that is, $(\mu - 370)/\sqrt{400} \sim t_8$ and so the (prior) probability that the rock will be older than 400 million years is

$$Pr(\mu > 400) = 0.0839$$

– calculated using the R command `1-pt((400-370)/sqrt(400), 8)`.

As before, these probability calculations demonstrate the benefit of taking the chemical measurements. Table 2.6 shows the probability that the rock is older than 400 million years under various scenarios: with and without the chemical measurements and assuming the variance of the age distribution is known or unknown. Notice that the

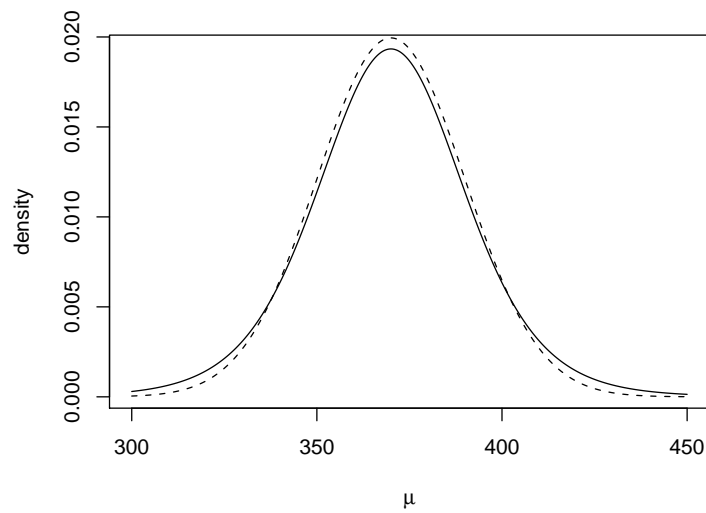


Figure 2.15: Marginal prior densities for μ : new version (solid) and previous version (dashed)

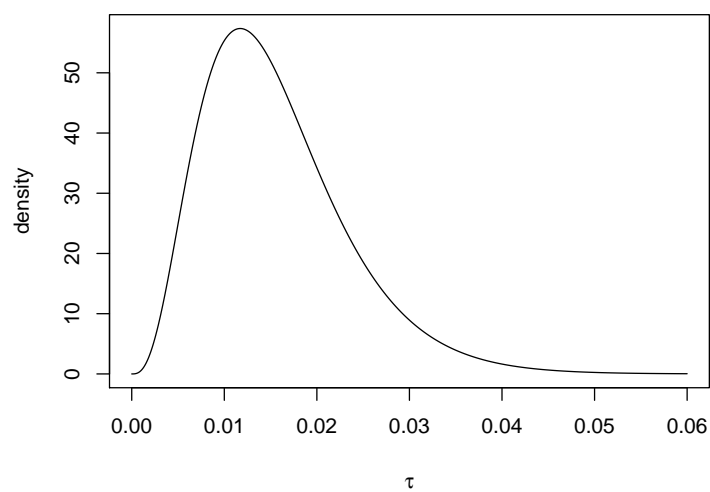


Figure 2.16: Marginal prior density for τ

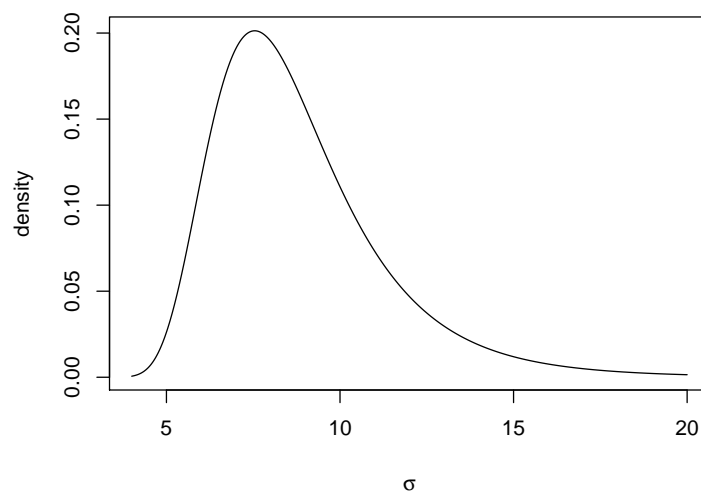


Figure 2.17: Marginal prior density for σ

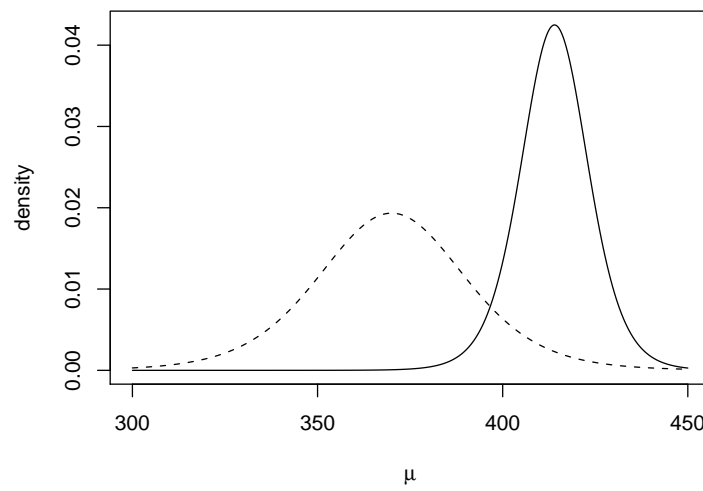


Figure 2.18: Prior (dashed) and posterior (solid) densities for the age of the rock

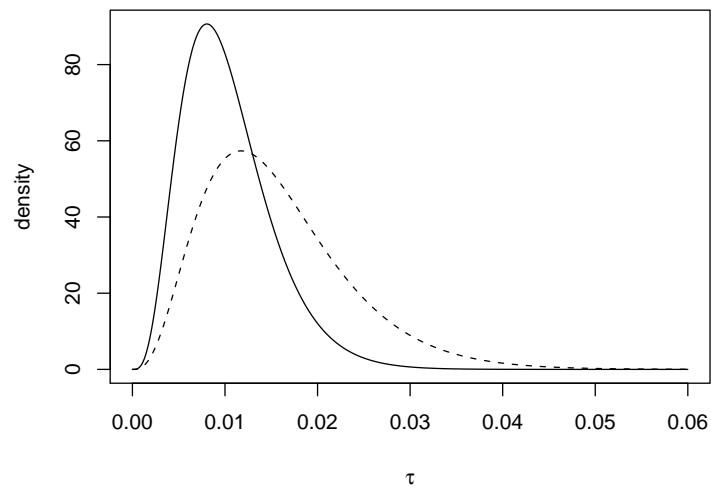


Figure 2.19: Prior (dashed) and posterior (solid) densities for τ

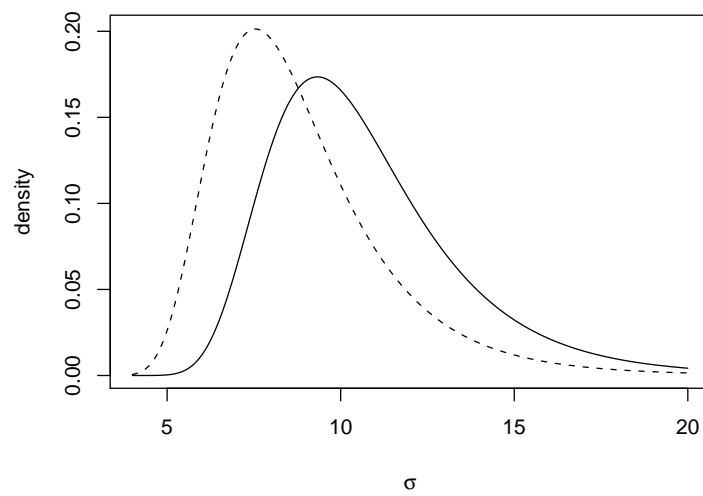
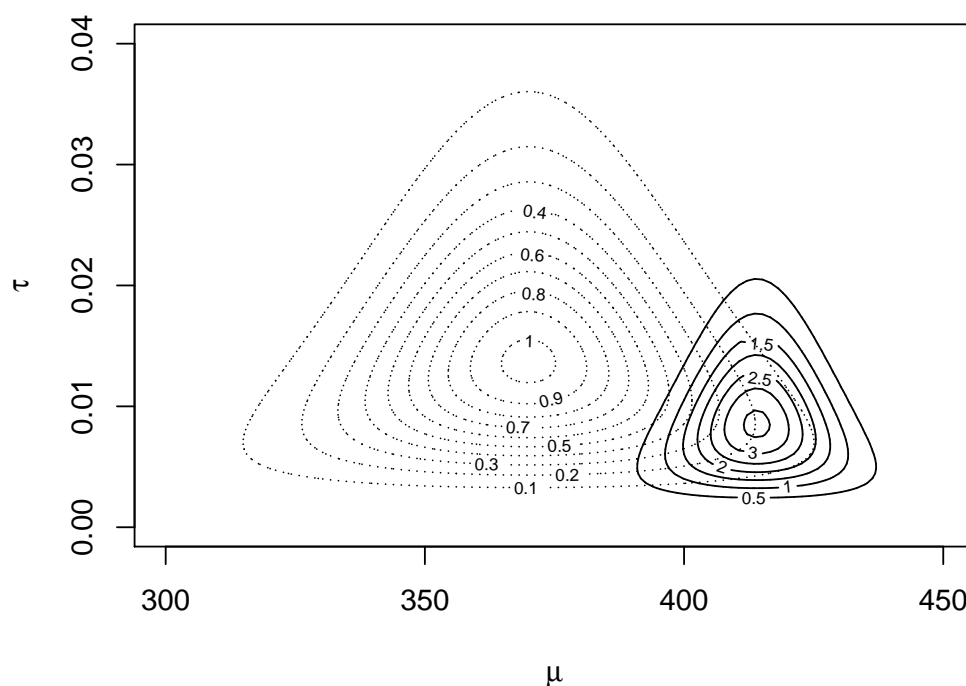


Figure 2.20: Prior (dashed) and posterior (solid) densities for σ

Figure 2.21: Contour plot of the prior (dashed) and posterior (solid) densities for (μ, τ)

	$Pr(\mu > 400)$	$Pr(\mu > 400 x = 421)$
Known variance	0.0668	0.9702
Unknown variance	0.0839	0.9202

Table 2.6: Probability calculations for the age of the rock

prior probability for the known variance model is (slightly) smaller than for the unknown variance model: difference = -0.0171 . However, the posterior probability for the known variance model is (slightly) larger than for the unknown variance model: difference = $+0.05$. This effect is common in Bayesian Statistics, *viz.* making stronger assumptions in a model will lead to more “confident” conclusions.