

# Chapter 3

## Priors

### 3.1 Introduction

In this chapter we will think about how to construct a suitable prior distribution  $\pi(\theta)$  for our parameter of interest  $\theta$ . For example, why did we use a  $Be(77, 5)$  distribution for  $\theta$  in the music expert example (page 31)? Why did we use a  $Be(2.5, 12)$  distribution for  $\theta$  in the example about the video game pirate (page 34)? And why did we assume a  $Ga(10, 4000)$  distribution for  $\theta$  in the earthquake example (page 37)? *Prior elicitation* – the process by which we attempt to construct the most suitable prior distribution for  $\theta$  – is a huge area of research in Bayesian Statistics; the aim in this course is to give a brief (and relatively simple) overview.

We will consider the case of *substantial prior knowledge*, where expert opinion, for example, gives us good reason to believe that some values in a permissible range for  $\theta$  are more likely to occur than others. In particular, we will re-visit the examples about the music expert, the video game pirate and earthquakes in Chapter 2.

Of course, sometimes it might be very difficult to properly elicit a prior distribution for  $\theta$ . For example, there may be no expert available to help guide your choice of distribution. In this case, the chosen prior distribution  $\pi(\theta)$  might be one which keeps the mathematics simple when operating Bayes' Theorem, whilst also assuming a large or “infinite” variance for  $\theta$  (*vague prior knowledge*). Alternatively, a prior which assumes all values of  $\theta$  are equally likely could be used to represent complete *prior ignorance*, as in the example about the possibly biased coin on page 28.

In this chapter we will also consider the construction of priors for  $\theta$  under certain parameter constraints, including the construction of *truncated priors*.

## 3.2 Substantial prior knowledge

In this section, we will consider the situation where we have substantial prior knowledge. In Examples 3.3 and 3.4 we will make use of some online software developed by researching Bayesian statisticians at Sheffield University: Professor Tony O’Hagan and Dr. Jeremy Oakley. We will use this software to implement the *trial roulette* and *bisection* methods of prior elicitation. This software is very simple to use and will be demonstrated in lectures; you will also be expected to use this in our next computer practical class.

### Example 3.1 (Using suggested prior summaries)

Let us return to Example 2.4 of Chapter 2. Recall that we were given some data on the “waiting times”, in days, between 21 earthquakes, and we discussed why an exponential distribution  $Exp(\theta)$  might be appropriate to model the waiting times. Further, we were told that an expert on earthquakes has prior beliefs about the rate  $\theta$ , described by a  $Ga(10, 4000)$  distribution; a plot of this prior is shown in Figure 2.8. Where did this prior distribution come from?

Suppose the expert tells us that earthquakes in the region we are interested in usually occur less than once per year; in fact, they occur on average once every 400 days. This gives us a rate of occurrence of about  $1/400 = 0.0025$  per day (to match the “daily” units given above). Further, he is fairly certain about this and specifies a very small variance of  $6.25 \times 10^{-7}$ .

A  $Ga(a, b)$  distribution seems sensible, since we can’t observe a negative daily earthquake rate and the Gamma distribution is specified over positive values only. Using the information provided by the expert, verify our use of  $a = 10$  and  $b = 4000$ .



...Solution to Example 3.1...

**Example 3.2 (Using suggested prior summaries)**

Now let us return to Example 2.2 of Chapter 2. We considered an experiment to determine how good a music expert is at distinguishing between pages from Haydn and Mozart scores; when presented with a score from each composer, the expert makes the correct choice with probability  $\theta$ .

Before conducting the experiment, we were told that the expert is very competent; in fact, we were told that  $\theta$  should have a prior distribution peaking at around 0.95 and for which  $\Pr(\theta < 0.8)$  is very small. To achieve this, we assumed that  $\theta \sim Be(77, 5)$ , with density given in Figure 2.4. How did we know a beta distribution would be appropriate? And how did we figure out the parameters of this distribution, i.e.  $a = 77$  and  $b = 5$ ?

By now, you should understand why we might work with a beta distribution: in this example,  $\theta$  is a probability and so must lie in the interval  $[0, 1]$ , and a beta distribution is defined over this range. But how did we know that  $a = 77$  and  $b = 5$  would give the desired properties for  $\theta$ ?

We are told that the mode of the distribution should be around 0.95; using the formulae on page 25, we can thus write

$$\frac{a - 1}{a + b - 2} = 0.95 \quad (3.1)$$

$$\begin{aligned} \Rightarrow a - 1 &= 0.95(a + b - 2) \\ \Rightarrow a - 0.95a &= 0.95b - 1.9 + 1 \\ \Rightarrow 0.05a &= 0.95b - 0.9 \\ \Rightarrow a &= 19b - 18. \end{aligned} \quad (3.2)$$

We are also told that  $\Pr(\theta < 0.8)$  must be small. In fact, suppose we are told that  $\theta < 0.8$  might occur with probability 0.0001. This means that if we integrate the probability density function for our beta distribution between 0 and 0.8, we would get 0.0001; from Equation (2.1) on page 25, we can write this as

$$\begin{aligned} \int_0^{0.8} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} d\theta &= 0.0001, \quad \text{i.e.} \\ \int_0^{0.8} \frac{\theta^{(19b-18)-1}(1-\theta)^{b-1}}{B(19b-18,b)} d\theta &= 0.0001, \end{aligned} \quad (3.3)$$

i.e. we set the cumulative distribution function for a  $Be(19b - 18, b)$ , evaluated at 0.8, equal to 0.0001 and solve for  $b$ . Although this would be rather tricky to do by hand, we can do this quite easily in R. Recall that the R command `dbeta(x,a,b)` evaluates the density of the  $Be(a,b)$  distribution at the point  $x$  (see page 25); the command `pbeta(x,a,b)` evaluates the corresponding cumulative distribution function at  $x$ . First of all, we re-write (3.3) to set it equal to zero:

$$\int_0^{0.8} \frac{\theta^{(19b-18)-1}(1-\theta)^{b-1}}{B(19b-18,b)} d\theta - 0.0001 = 0. \quad (3.4)$$

We then write a function in R which computes the left-hand-side of Equation (3.4):

```
f=function(b){
+  answer=pbeta(0.8,((19*b)-18),b)-0.0001
+  return(answer)}
```

The trick now is to use a numerical procedure to find the root of `answer` in the R function above, i.e. find the value `b` which equates `answer` to zero (as is required in Equation (3.4)). We can do this using the R function `uniroot(f, lower=, upper=)`, which uses a numerical search algorithm to find the root of the expression provided by the function `f`, having been given a `lower` bound and an `upper` bound to search within. We know from the formulae on page 25 that  $a, b > 1$  when using expression (3.1) for the mode, so we can search for a root over some specified domain  $> 1$ : for example, we might use `lower=1` and `upper=100`, giving:

```
> uniroot(f,lower=1,upper=100)
$root
[1] 5.06513

$f.root
[1] 6.008134e-09

$iter
[1] 14

$estim.prec
[1] 6.103516e-05
```

Thus, the solution to Equation (3.3) is  $b = 5.06513$ . For simplicity, rounding down to  $b = 5$  and then substituting into (3.2) gives

$$a = 19 \times 5 - 18 = 77,$$

hence the use of  $\theta \sim Be(77, 5)$  in Example 2.2 in Chapter 2.

In the next two examples, we will use the *MATCH Uncertainty Elicitation Tool*, developed by Professor Tony O'Hagan and Dr. Jeremy Oakley at Sheffield, to demonstrate two techniques of prior elicitation: the *trial roulette method* and the *bisection method*.

### Example 3.3 (Trial roulette method)

We now return to Example 2.3 in Chapter 2. Recall that Max is a video game pirate, and he is trying to identify the proportion  $\theta$  of potential customers who might be interested in buying *Call of Duty: Elite* next month. Why did we use  $\theta \sim Be(2.5, 12)$ ?

For each month over the last two years Max knows the proportion of his customers who have bought similar games; these proportions are shown below in Table 3.1.

0.32	0.25	0.28	0.15	0.33	0.12	0.14	0.18	0.12	0.05	0.25	0.08
0.07	0.16	0.24	0.38	0.18	0.15	0.22	0.05	0.01	0.19	0.08	0.15

Table 3.1: Proportion of Max’s customers, each month, who have bought computer games in the same genre as *Call of Duty: Elite*.

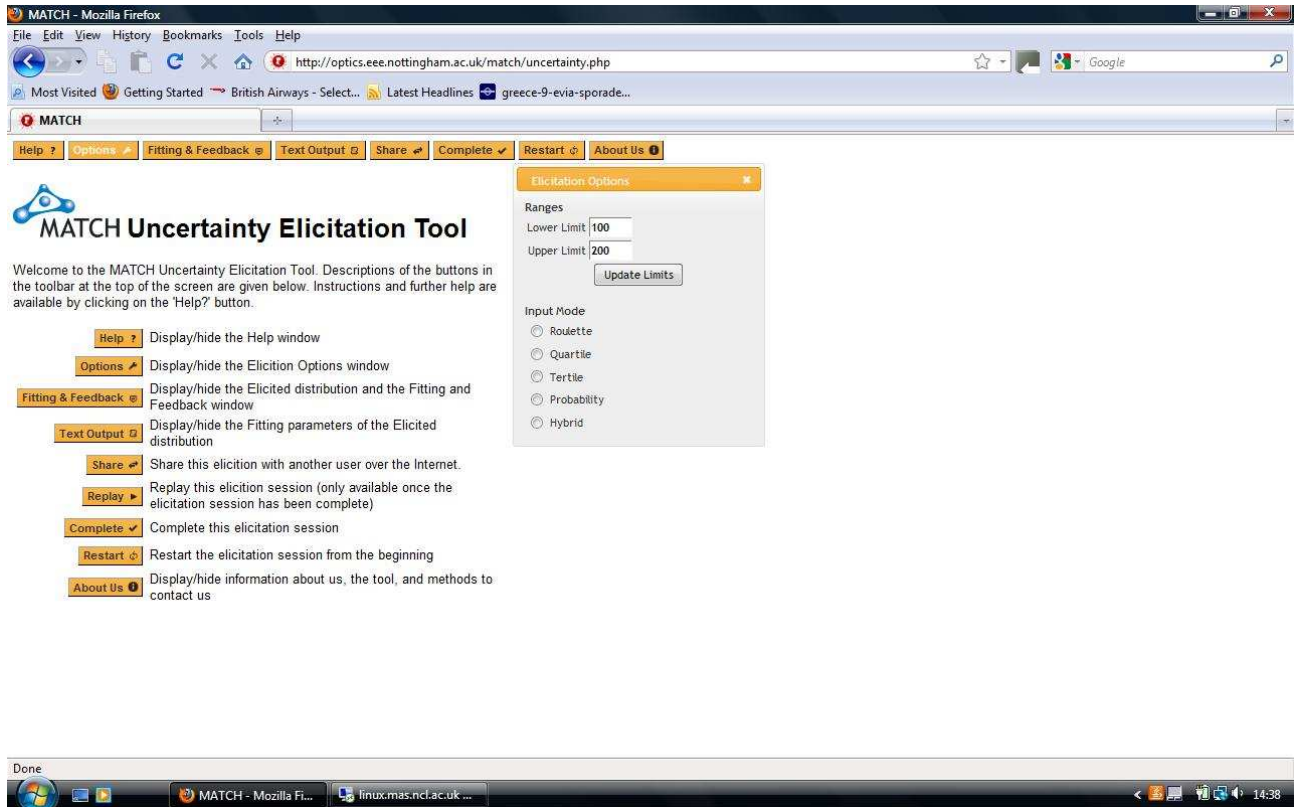
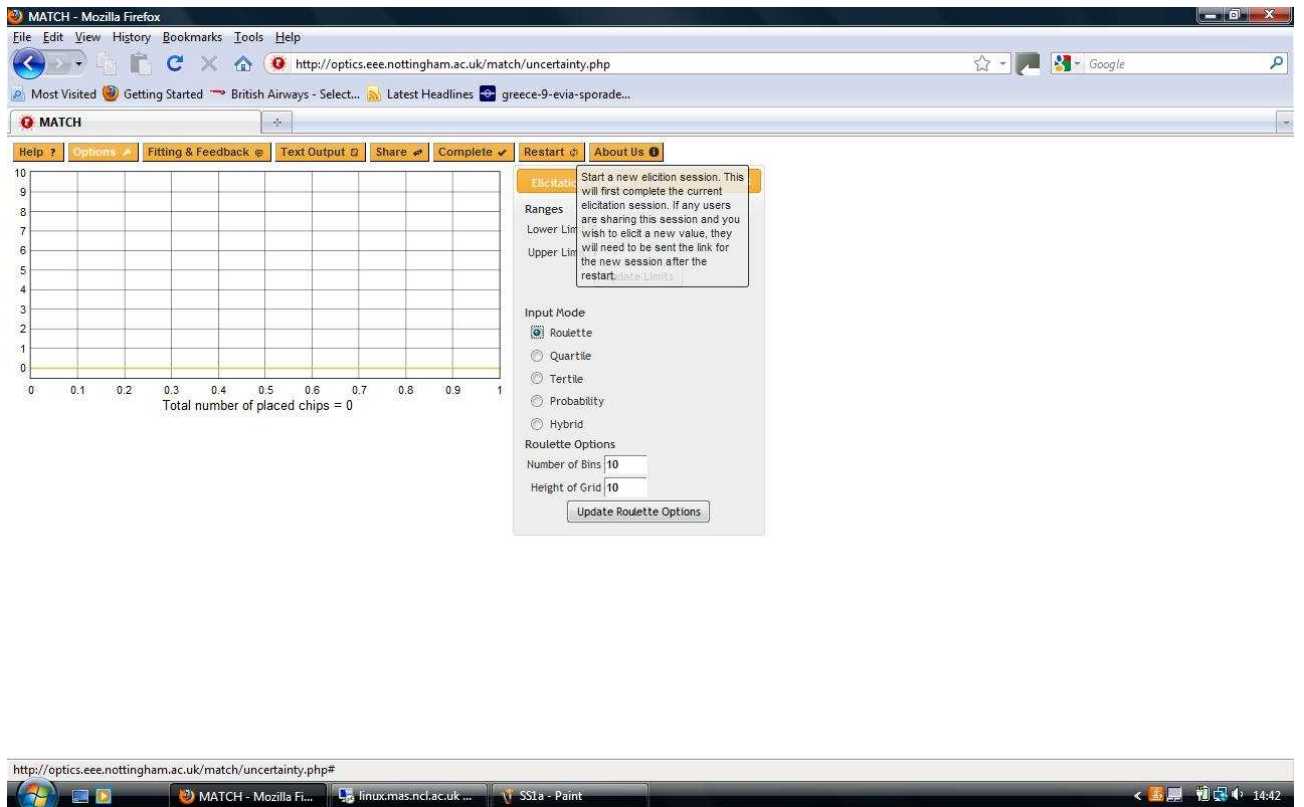
We will use these data to form a prior for  $\theta$  using the *trial roulette method*. Here, we divide the sample space for  $\theta$  into  $m$  bins, and we distribute  $n$  chips amongst the bins, with the proportion of chips allocated to a particular bin representing the probability of  $\theta$  lying in that bin. If this is done graphically, we can see the shape of our distribution forming as we allocate the chips. Once the allocation is done, we can then fit a parametric distribution to match the shape of the distribution of chips. Actually, we will do this using the *MATCH Uncertainty Elicitation Tool*. You can enter *MATCH* directly by typing the following URL into any web browser:

<http://optics.eee.nottingham.ac.uk/match/uncertainty.php>

Doing so gives the screenshot shown in Figure 3.1 overleaf. Notice the default **Lower Limit** and **Upper Limit** are set at 100 and 200, respectively; in this example,  $\theta$  is a proportion, and so these need to be changed to 0 and 1 respectively. Doing so, and then selecting the **Roulette** option, gives what can be seen in the screenshot in Figure 3.2. The **Number of Bins** and **Height of Grid** can be changed, if necessary. We can now place chips in the bins according to the prior information given in Table 3.1. For example, the first chip will go in between 0.3 and 0.4 as the first proportion is 0.32. Completing the grid gives the distribution of chips as shown in Figure 3.3.

Finally, selecting the **Fitting and Feedback** button gives the screenshot shown in Figure 3.4. Here, you can see that *MATCH* automatically chooses the best-fitting distribution – in this case, a **Scaled Beta** distribution (in this example just the beta distribution that you should be familiar with by now); the program also returns a picture of the fitted density function, showing the parameters of this distribution underneath. Here, *MATCH* suggests using  $a \approx 2.5$  and  $b \approx 12$  (given as  $\alpha$  and  $\beta$  on the screen); hence, the use of  $\theta \sim Be(2.5, 12)$  in Example 2.3 in Chapter 2.

Of course, instead of using past data to complete the roulette grid, we could ask an expert to place chips in the bins according to her beliefs about likely values of  $\theta$ ; twice as many chips in one bin compared to another would mean that the expert believes that  $\theta$  is twice as likely to take on values in that bin than in the other. Note the **Feedback Percentiles** in Figure 3.4; we will discuss their use in the next example.

Figure 3.1: The homepage of the *MATCH Uncertainty Elicitation Tool*.Figure 3.2: The *Roulette* option in *MATCH*.

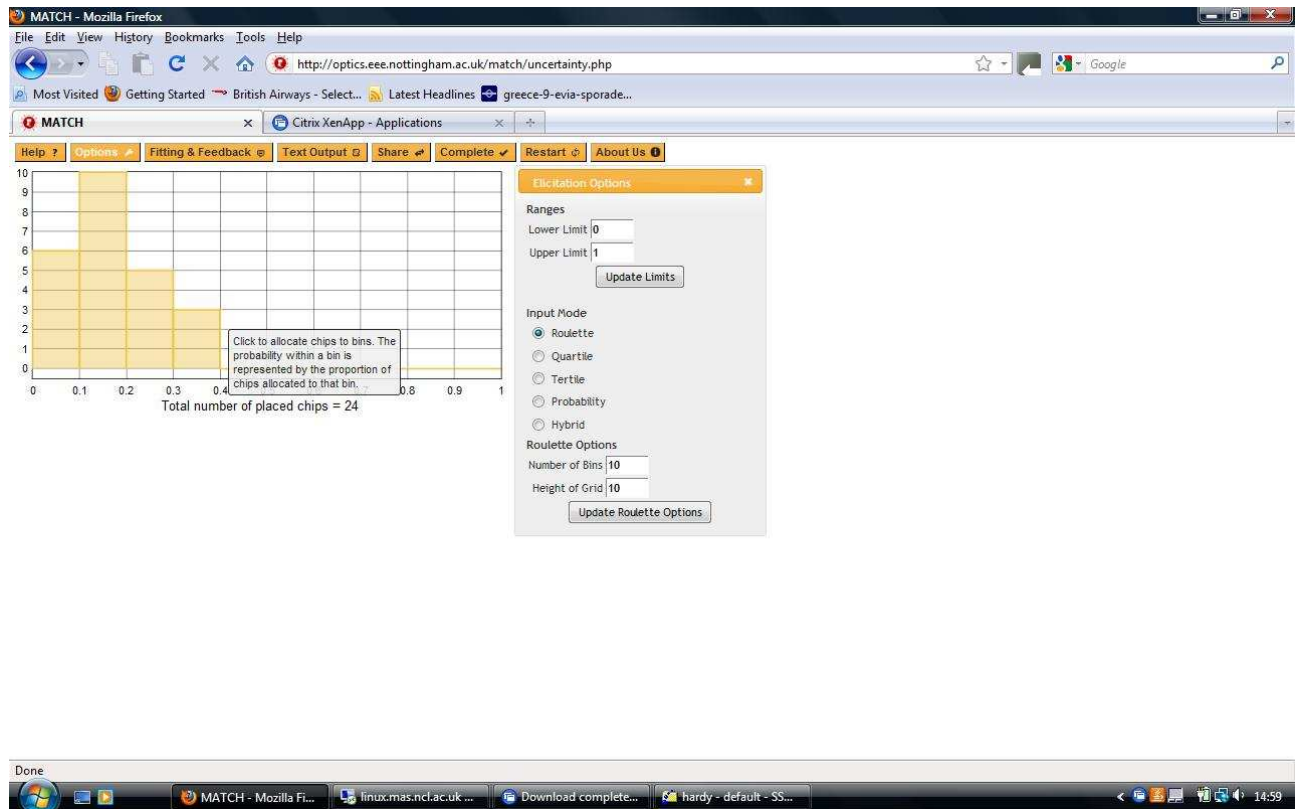


Figure 3.3: Placing chips in bins using the *Roulette* option in *MATCH*.

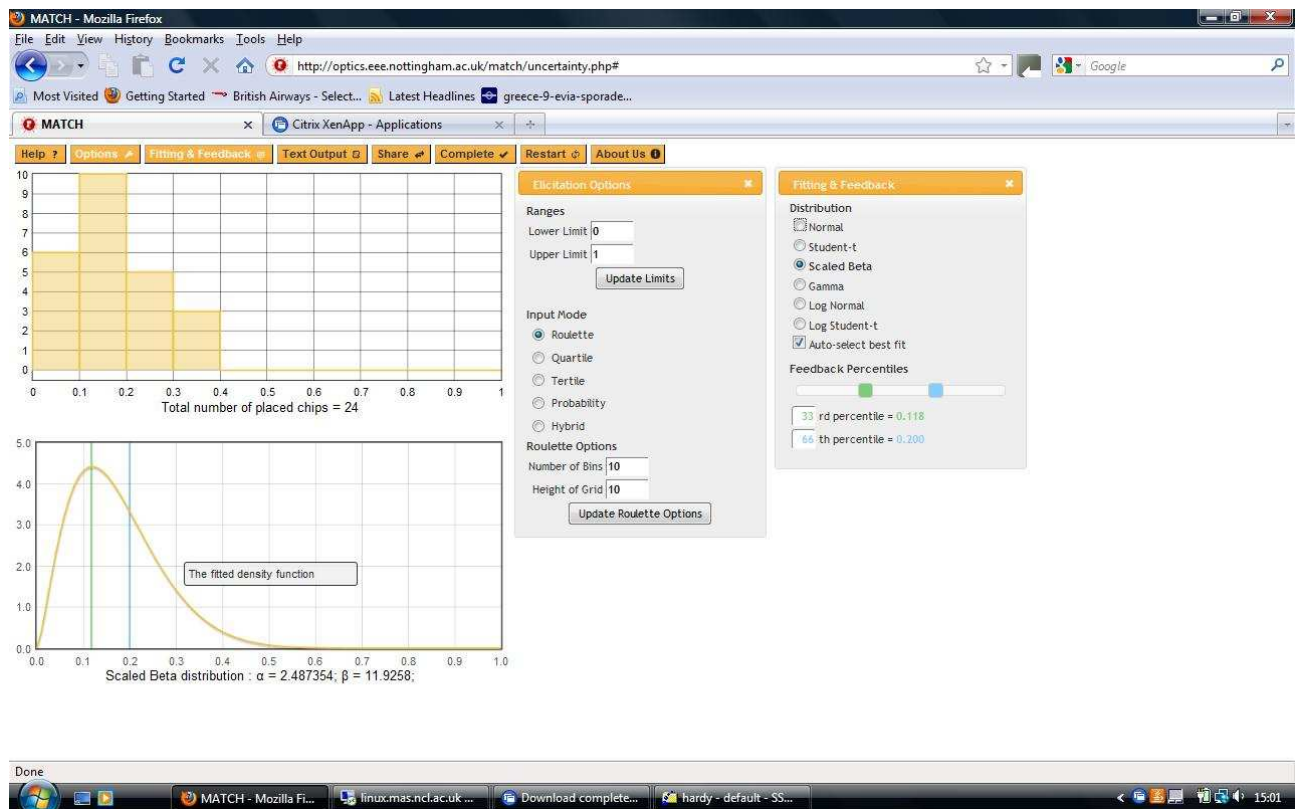


Figure 3.4: Fitting and Feedback in *MATCH*.

### Example 3.4 (Bisection method)

The *bisection method* is another method of prior elicitation. This method requires the expert to be able to split the range of allowable values for  $\theta$  into four smaller, equally likely, chunks.

First of all, the range is bisected into two intervals of equal probability from which we elicit the expert's median,  $m$ ; we then ask the expert to further bisect each of these intervals to elicit the lower and upper quartiles ( $l$  and  $u$ , respectively). From these elicited quartiles, we can then obtain a parametric distribution for  $\theta$ ; for this we will again make use of *MATCH*. Finally, a stage of feedback and refinement is considered before we settle on a distribution for  $\theta$ ; this stage of feedback and refinement can (and should!) be undertaken in all situations where we attempt to elicit a prior distribution from information given to use by an expert. We will now demonstrate the bisection method through an example.

Over the past 15 years there has been considerable scientific interest in the rate of retreat,  $\theta$  (feet 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. We are interested in eliciting a suitable prior distribution for  $\theta$  for the *Zachariae Isstrøm* glacier in Greenland.

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  $\theta$ , respectively. We now attempt to elicit the median and quartiles for  $\theta$  from the glaciologist.

#### 1. Elicit the expert's median

*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 an 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. The value which bisects the whole range, in terms of how likely certain values of  $\theta$  are to occur, should be closer to the lower bound than the upper.”

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 3.5. Notice that the green area of the median bar represents the lower 50% of the distribution for  $\theta$  and the blue area represents the upper 50% of the distribution for  $\theta$ . Notice also that the lower and upper quartiles are still at their default values – one quarter and three quarters of the range, respectively.



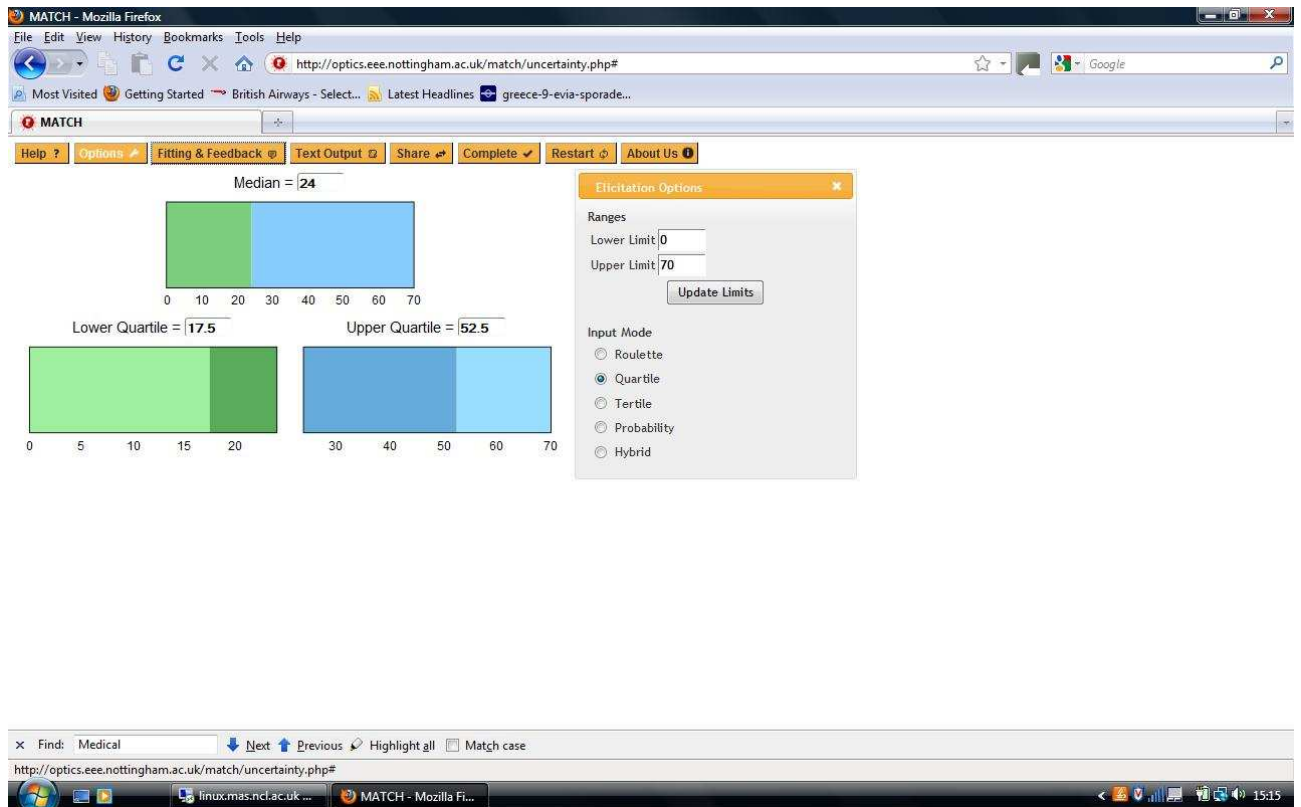


Figure 3.5: Entering the lower and upper limits, and the elicited median, in *MATCH*.

## 2. Elicit the expert's lower quartile

*Statistician*: “So, you think there is an evens 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

*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 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 3.6 shows a screenshot of this.

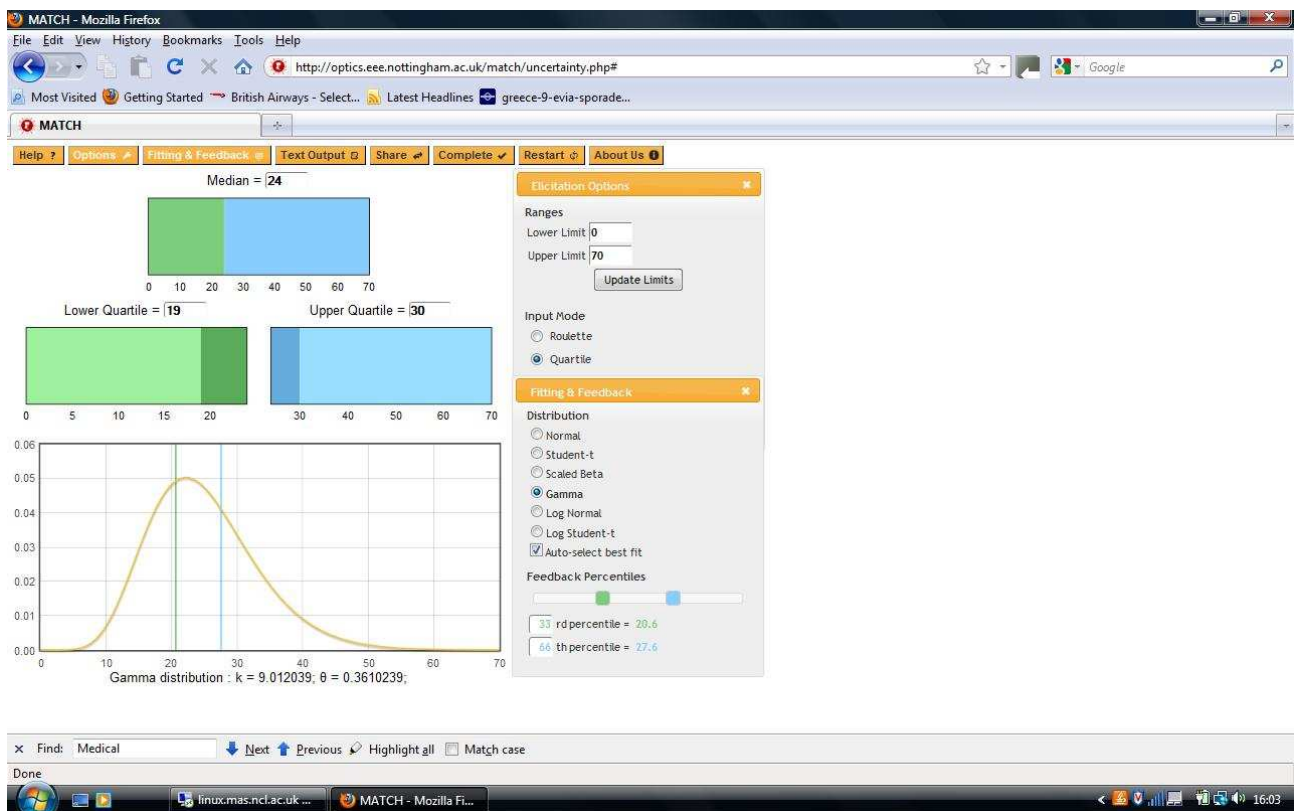


Figure 3.6: Entering the lower and upper quartiles to obtain a fitted distribution for  $\theta$  in *MATCH*.

#### 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 3.6, indicates that a  $Ga(9, 0.36)$  distribution might be appropriate for  $\theta$ . Notice from Figure 3.6 that *MATCH* also gives **Feedback percentiles** – we will consider these in the feedback and refinement stage.

#### 6. Feedback and refinement

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

*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 3.6. Do you think this reasonably specifies your beliefs about the retreat of the *Zachariae Isstrøm* glacier this year?”

*Glaciologist:* “I think this looks OK. The plot peaks close to my suggested value of about 24 feet; to allows for all values between 0 and 70 feet, but gives diminishing probabilities as we move towards these extremes.”

The Statistician should now consider the tails: we can do this by setting the **Feedback Percentiles** in *MATCH* to 1% and 99% (see Figure 3.7).

*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.”

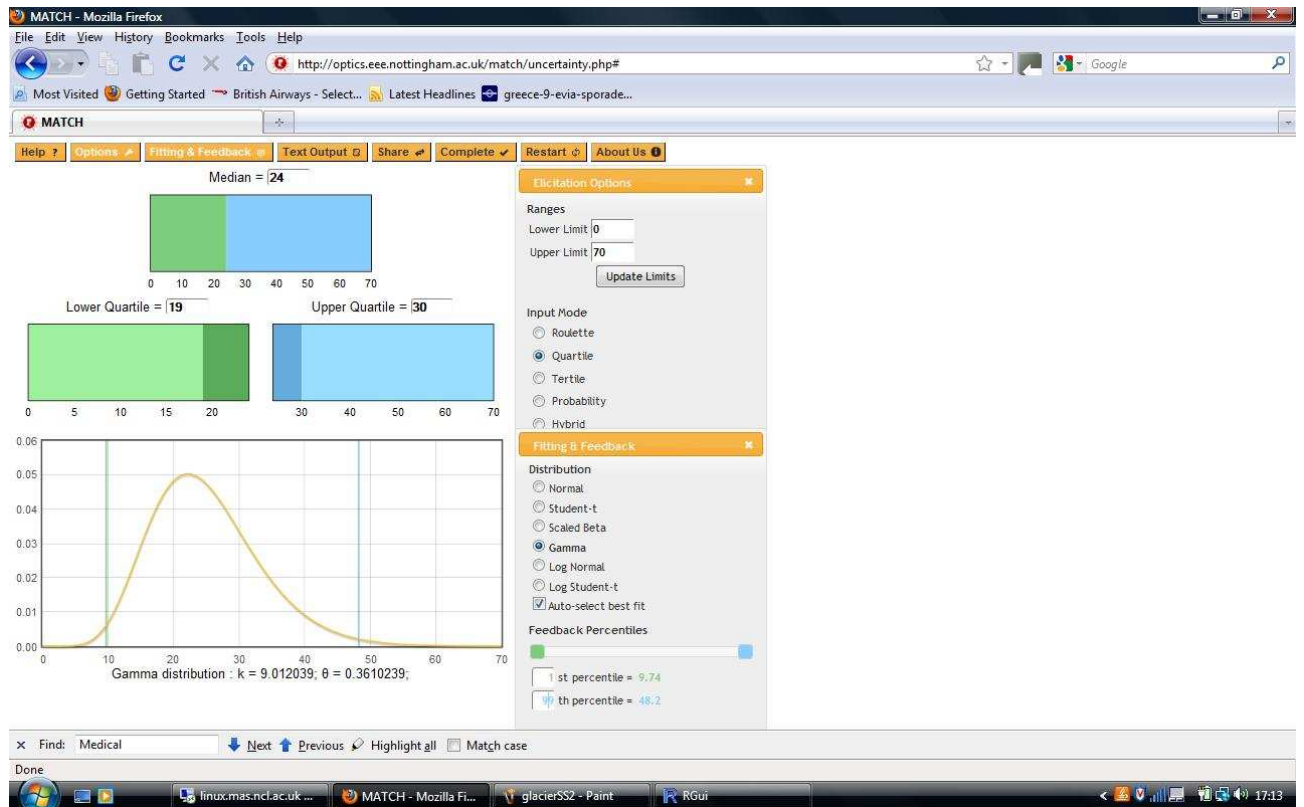



Figure 3.7: Obtaining tail probabilities in *MATCH*: here, we ask for the 1st percentile and the 99th percentile.

### Example 3.5


Let  $Y$  be the retreat, in feet, of the *Zachariae Isstrøm* glacier. A *Pareto* distribution with rate  $\theta$  is often used to model such geophysical activity, with probability density function

$$f(y|\kappa, \theta) = \theta \kappa^\theta y^{-(\theta+1)}, \quad \theta, \kappa > 0 \text{ and } y > \kappa.$$


- (a) Obtain the likelihood function for  $\theta$  given the parameter  $\kappa$  and some observed data  $y_1, y_2, \dots, y_n$  (independent).

 ...Solution to Example 3.5(a)...

- (b) Suppose we observe a retreat of 20 feet at the *Zachariae Isstrøm* glacier in 2012. Write down the likelihood function for  $\theta$ .

 ...Solution to Example 3.5(b)...

- (c) Using the elicited prior for the rate of retreat we obtained from the expert glaciologist in Example 3.4, and assuming  $\kappa$  is known to be 12, obtain the posterior distribution  $\pi(\theta|y_1 = 20)$ .

 ...Solution to Example 3.5(c)...

 ...Solution to Example 3.5(c) continued...

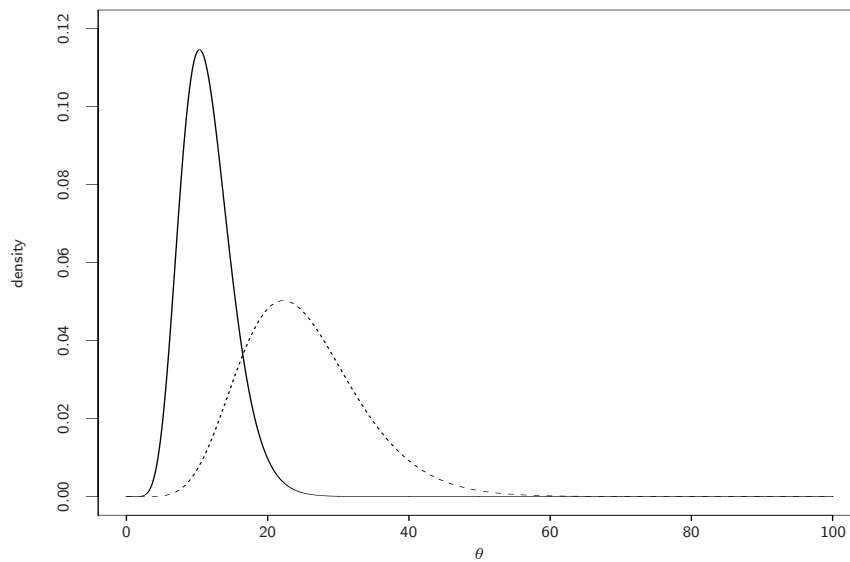


Figure 3.8: Prior (dashed) and posterior (solid) densities for the rate of glacial retreat at the *Zachariae Isstrøm* glacier.

**Definition 3.1**

We have *substantial prior information* for  $\theta$  when the prior distribution *dominates* the posterior distribution, that is  $\pi(\theta|\mathbf{x}) \sim \pi(\theta)$ . An example of substantial prior knowledge was given in Example 2.2 where a music expert was trying to distinguish between pages from Mozart and Haydn scores. Figure 3.9 shows the prior and posterior distributions for  $\theta$ , the probability that the expert makes the correct choice. Notice the similarity between the prior and posterior distributions. Observing the data has not altered our beliefs about  $\theta$  very much.

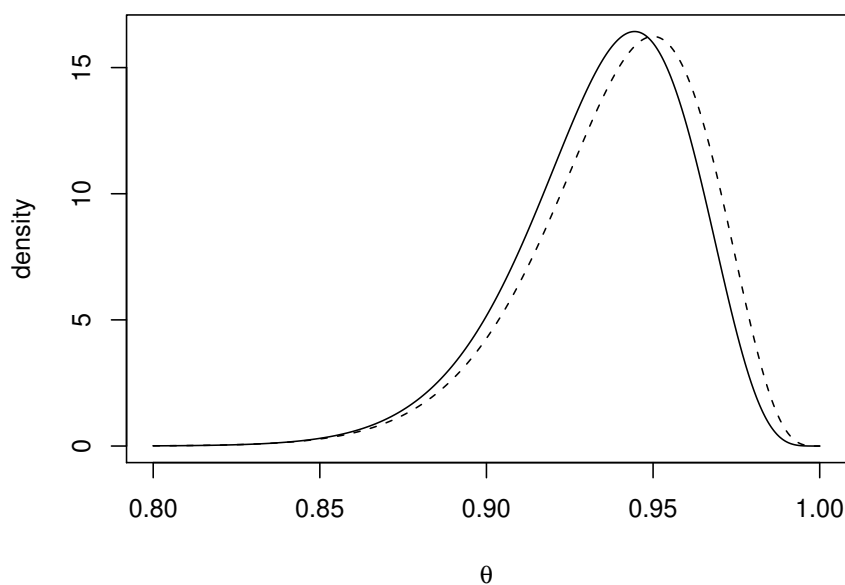


Figure 3.9: Prior (dashed) and posterior (solid) densities for  $\theta$

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

1. the intractability of the mathematics in deriving the posterior distribution — though with modern computing facilities this is less of a problem,
2. the practical formulation of the prior distribution — coherently specifying prior beliefs in the form of a probability distribution is far from straightforward although, as we have seen, this can be attempted using computer software.

### 3.3 Parameter constraints

Our prior information may include *parameter constraints*. For example an expert may tell us that it is physically impossible for a parameter to take values in a particular range. Often this translates mathematically to a condition such as  $\theta \geq 1$ . One way to define priors which satisfy such constraints is by using *truncated distributions*.

#### Definition 3.2

Consider a univariate continuous distribution with density  $\pi(\theta)$ . Then this distribution *truncated to the interval*  $[a, b]$  has density:

$$\pi_T(\theta) = \begin{cases} \pi(\theta)/k & \text{for } \theta \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

where  $k = \int_a^b \pi(\theta) d\theta$ .

Notes:

1. For one-side constraints – e.g.  $\theta \geq 1$  – we can take  $a = -\infty$  or  $b = \infty$ .
2. Strict inequalities such as  $\theta > 1$  can be treated in exactly the same way as non-strict inequalities such as  $\theta \geq 1$  (this is because  $\theta$  is a continuous random variable).
3. For Bayesian analysis we usually don't need to actually calculate  $k$ , as we can simply use:

$$\pi_T(\theta) \propto \begin{cases} \pi(\theta) & \text{for } \theta \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

#### Example 3.6

Suppose  $\theta \sim N(b, d^2)$ . Find the density of the truncated distribution for  $\theta > 0$ .



...Solution to Example 3.6...



 ...Solution to Example 3.6 continued...

Figure 3.10 shows a plot of the densities of (a) a  $N(1, 1)$  distribution and (b) a  $N(1, 1)$  distribution truncated to  $\theta > 0$ . Notice that for  $\theta > 0$  the truncated normal's density has the same shape as that of the original distribution. However the truncated density takes proportionately larger values, since both curves must have an area underneath of one. Also, we can see that the mean of the truncated distribution will be larger than the mean of the original distribution and the standard deviation will be smaller.

This is an important general point. Truncating a distribution changes the mean and variance. Calculating the new values can be difficult.

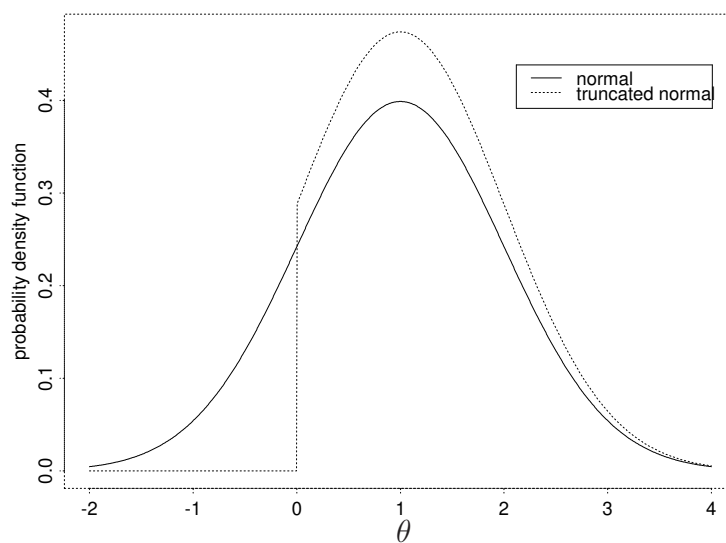


Figure 3.10: Plot of a normal and truncated normal distribution

## Bayesian statistics with truncated priors

Suppose we can do a Bayesian analysis for an ordinary prior. Then it's easy to do the analysis for a truncated version of this prior by the following theorem.

### Theorem: Truncated posteriors

Suppose that for a prior  $\pi(\theta)$  the resulting posterior is  $\pi(\theta|\mathbf{x})$ . Let  $\pi_T(\theta)$  be the result of truncating the prior to  $[a, b]$ . Then the corresponding posterior is  $\pi(\theta|\mathbf{x})$  truncated to  $[a, b]$ .

### Proof

Let  $\pi'(\theta|\mathbf{x})$  be the posterior for the truncated prior. Then:

$$\begin{aligned}\pi'(\theta|\mathbf{x}) &\propto \pi_T(\theta)f(\mathbf{x}|\theta) \\ &\propto \begin{cases} \pi(\theta)f(\mathbf{x}|\theta) & \text{for } \theta \in [a, b] \\ 0 & \text{otherwise} \end{cases} \\ &\propto \begin{cases} \pi(\theta|\mathbf{x}) & \text{for } \theta \in [a, b] \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

This result makes it easy to do Bayesian analysis under parameter constraints.

### Example 3.7

Consider again the case of  $X_i|\theta \sim N(\theta, h^2)$ ,  $i = 1, 2, \dots, n$  (independent) and  $\theta \sim N(b, d^2)$ , with  $h$  known. Suppose we knew in advance that the experiment could only result in positive values for  $\theta$ . Find the posterior distribution for  $\theta$ .



...Solution to Example 3.7...

 ...Solution to Example 3.7 continued...

Figure 3.11 plots the  $N(1, 1)$  posterior densities using the original and truncated priors. This plot highlights an important consequence of using truncated distributions for modelling prior beliefs, namely, that if particular parameter values are ruled out prior to seeing the data then they are also ruled out after seeing the data. Normally, this is not a problem but, as the following example shows, when the truncation in a prior distribution does not include parameter values for which the likelihood function is large, misleading conclusions can be made.

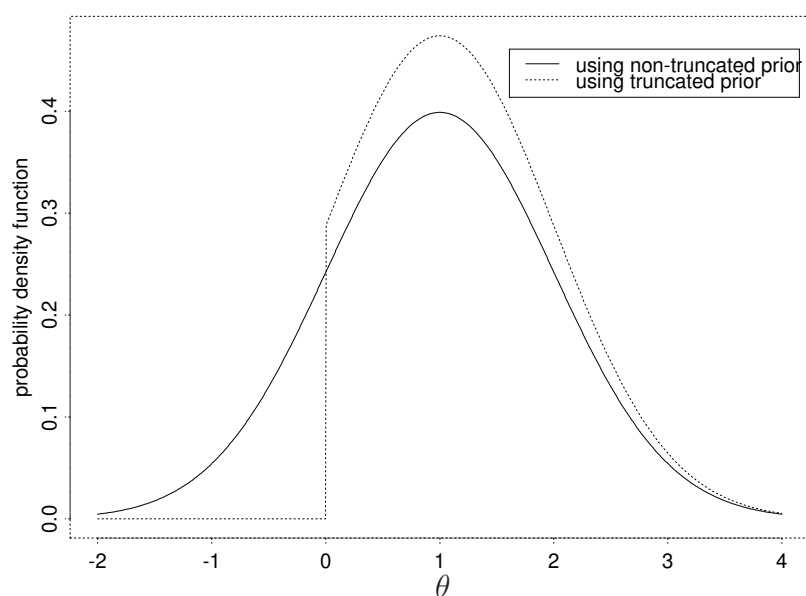


Figure 3.11: Plot of a posterior distribution determined using a non-truncated and a truncated prior distribution

### Example 3.8

Consider the case of a large random sample ( $n = 1500$ ) with sample mean  $\bar{x} = 157/15 \simeq 10.47$  from a normal distribution with known variance ( $h^2 = 100$ ) and a normal  $N(3, 1)$  prior distribution for the mean parameter  $\theta$ . The prior probability that  $\theta$  exceeds 9 is almost zero:  $\Pr(\theta > 9) \simeq 10^{-9}$ . So what is the effect of using a prior distribution which rules out the extremely unlikely values of  $\theta$  greater than 9? Put another way, is there much difference between the posterior distributions calculated using the prior with  $\Pr(\theta > 9) \simeq 10^{-9}$  or a truncated version of the prior with  $\Pr(\theta > 9) = 0$ ?

If no truncation is applied to the prior distribution then Bayes Theorem produces the posterior distribution  $\theta|\mathbf{x} \sim N(10, 0.25^2)$ . The discussion preceding this example tells us that imposing the truncation  $\theta < 9$  on the prior distribution produces a posterior distribution which is a  $N(10, 0.25^2)$  distribution, truncated to  $\theta < 9$ . Figure 3.12 shows the resulting posterior densities. Clearly, truncating the prior distribution to  $\theta < 9$  has resulted in a posterior distribution truncated to  $\theta < 9$ , even though the likelihood function is very peaked at  $\theta \simeq 10.47$ . So our prior has ruled out the most likely values according to the data! Using a prior distribution which gives very small – but non-zero – probability to values of  $\theta > 9$ , avoids this problem.

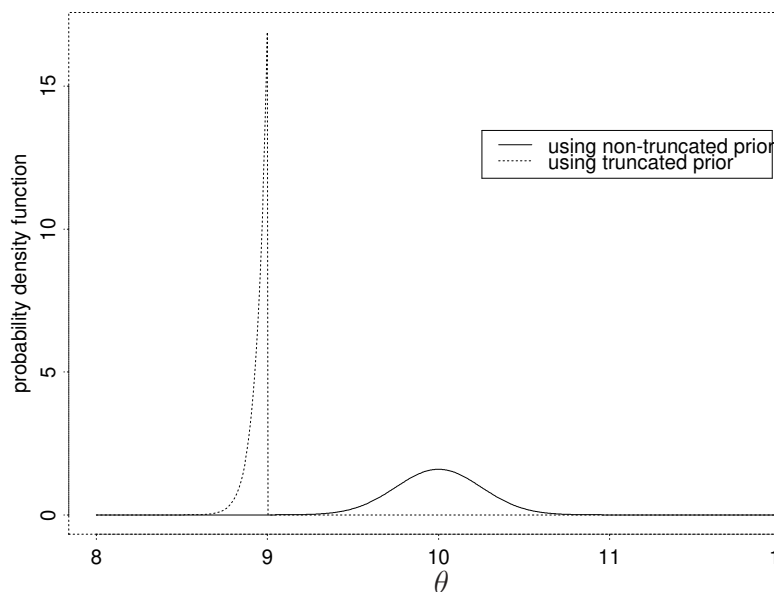


Figure 3.12: Plot of a posterior distribution determined using a non-truncated and a truncated prior distribution

This example motivates the pragmatic rule: never rule out values for parameters which are very implausible but not impossible. Instead these parameter values should be given very small probability density. The data will then be allowed to inform the posterior distribution about values of  $\theta$  with very low prior probability (density) but with very high likelihood.

*“If a decision-maker thinks something cannot be true and interprets this to mean it has zero probability, he will never be influenced by any data, which is surely absurd. So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved” – Dennis Lindley*

### 3.4 Vague Prior Knowledge/Prior Ignorance

If we have very little or no prior information about the model parameters  $\theta$ , we must still choose a prior distribution in order to operate Bayes Theorem. Obviously, it would be sensible to choose a prior distribution which is not concentrated about any particular value, that is, one with a very large variance. In particular, most of the information about  $\theta$  will be passed through to the posterior distribution via the data, and so we have  $\pi(\theta|\mathbf{x}) \sim f(\mathbf{x}|\theta)$ .

An example of vague prior knowledge was given in Example 2.1 where a possibly biased coin was assessed. Figure 3.13 shows the prior and posterior distributions for  $\theta = \text{Pr}(\text{Head})$ . Notice that the prior and posterior distributions look very different. In fact, in this example, the posterior distribution is simply a scaled version of the likelihood function – likelihood functions are not usually proper probability (density) functions and so scaling is required to ensure that it integrates to one. Most of our beliefs about  $\theta$  have come from observing the data.

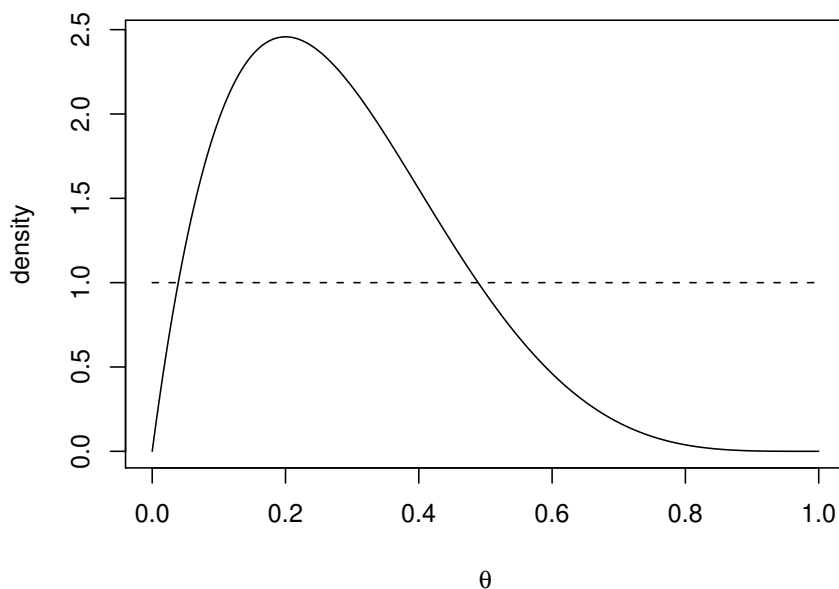


Figure 3.13: Prior (dashed) and posterior (solid) densities for  $\theta$

## Vague Prior Knowledge

We represent vague prior knowledge by using a prior distribution which is conjugate to the model for  $\boldsymbol{x}$  and which has “infinite” variance.

### Example 3.9


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



*...Solution to Example 3.9...*

**Example 3.10**

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

 *...Solution to Example 3.10...***Prior Ignorance**

We could represent ignorance by the concept “all values of  $\theta$  are equally likely”. If  $\theta$  were discrete with  $m$  possible values then we could assign each value the same probability  $1/m$ . However, if  $\theta$  is continuous, we need some limiting argument (from the discrete case). Suppose that  $\theta$  can take values between  $a$  and  $b$ , where  $-\infty < a < b < \infty$ . Letting all (permitted) values of  $\theta$  be equally likely results in taking a uniform  $U(a, b)$  distribution as our prior distribution for  $\theta$ . However, if the parameter space is not finite then we cannot do this: there is no such thing as a  $U(-\infty, \infty)$  distribution. Convention

suggests that we should use the “improper” uniform prior distribution

$$\pi(\theta) = \text{constant}.$$

This distribution is improper because  $\int_{-\infty}^{\infty} \pi(\theta) d\theta$  is not a convergent integral, let alone equal to one. We have a similar problem if  $\theta$  takes positive values — we cannot use a  $U(0, \infty)$  prior distribution. Now if  $\theta \in (0, \infty)$  then  $\phi = \log \theta \in (-\infty, \infty)$ , and so we could use an “improper” uniform prior for  $\phi$ :  $\pi(\phi) = \text{constant}$ . In turn, this induces a distribution on  $\theta$ . Recall the result from Distribution Theory:

Suppose that  $X$  is a random variable with probability density function  $f_X(x)$ . If  $g$  is a bijective (1–1) function then the random variable  $Y = g(X)$  has probability density function

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|. \quad (3.7)$$

Applying this result to  $\theta = e^\phi$  gives

$$\begin{aligned} \pi_\theta(\theta) &= \pi_\phi(\log \theta) \left| \frac{d}{d\theta} \log \theta \right|, & \theta > 0 \\ &= \text{constant} \times \left| \frac{1}{\theta} \right|, & \theta > 0 \\ &\propto \frac{1}{\theta}, & \theta > 0. \end{aligned}$$

This too is an improper distribution.

There is a drawback of using uniform or improper priors to represent prior ignorance: if we are “ignorant” about  $\theta$  then we are also “ignorant” about any function of  $\theta$ , for example, about  $\phi_1 = \theta^3$ ,  $\phi_2 = e^\theta$ ,  $\phi_3 = 1/\theta$ , ... . Is it possible to choose a distribution where we are ignorant about all these functions of  $\theta$ ? If not, on which function of  $\theta$  should we place the uniform/improper prior distribution? After a little thought, it should be clear that there is no distribution which can represent ignorance for all functions of  $\theta$ . The above example shows that assigning a uniform/ignorance prior to  $\theta$  means that we do not have a uniform/ignorance prior for  $e^\theta$ .

A solution to problems of this type was suggested by Sir Harold Jeffreys. His suggestion was specified in terms of Fisher’s Information:

$$I(\theta) = E_{\mathbf{X}|\theta} \left[ -\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{X}|\theta) \right]. \quad (3.8)$$

He recommended that we represent prior ignorance by the prior distribution

$$\pi(\theta) \propto \sqrt{I(\theta)}. \quad (3.9)$$

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



**Example 3.11**

Suppose we have a random sample from a distribution with probability density function

$$f(x|\theta) = \frac{2x e^{-x^2/\theta}}{\theta}, \quad x > 0, \theta > 0.$$

Determine the Jeffreys prior for this model.




*...Solution to Example 3.11...*

 *...Solution to Example 3.11 continued...*

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

**Example 3.12**

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

 *...Solution to Example 3.12...*

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


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

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

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

**Example 3.13**

Suppose we have a random sample from a  $N(\mu, 1/\tau)$  distribution (with  $\tau$  known). Determine the Jeffreys prior for this model.

 *...Solution to Example 3.13...*

Notice that this distribution is improper since  $\int_{-\infty}^{\infty} d\mu$  is a divergent integral, and so we cannot find a constant which ensures that the density function integrates to one.

Also it is a limiting form of a  $N(b, 1/d)$  density (ignoring the integration constant) since

$$\left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \propto \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \rightarrow 1, \quad \text{as } d \rightarrow 0.$$

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

### 3.5 Asymptotic posterior distribution

There are many limiting results in Statistics. The one you will probably remember is the Central Limit Theorem. This concerns the distribution of  $\bar{X}_n$ , the mean of  $n$  independent and identically distributed random variables (each with known mean  $\mu$  and known variance  $\sigma^2$ ), as the sample size  $n \rightarrow \infty$ . It is easy to show that  $E(\bar{X}_n) = \mu$  and  $Var(\bar{X}_n) = \sigma^2/n$ , and so

$$E\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right] = 0 \quad \text{and} \quad Var\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right] = 1.$$

These two equations are true for all values of  $n$ . The important part of the Central Limit Theorem is the description of the distribution of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  as  $n \rightarrow \infty$ :

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} N(0, 1) \quad \text{as } n \rightarrow \infty.$$

The following theorem gives a similar result for the posterior distribution.

#### Theorem (Asymptotic posterior)

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

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

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

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

**Proof**

Using Bayes Theorem, the posterior distribution for  $\theta$  is

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

Let  $\psi = \sqrt{n}(\theta - \hat{\theta})$  and

$$\ell_n(\theta) = \frac{1}{n} \log f(\mathbf{x}|\theta)$$

be the average log-likelihood per observation, in which case,  $f(\mathbf{x}|\theta) = e^{n\ell_n(\theta)}$ . Using (3.7), the posterior distribution of  $\psi$  is

$$\begin{aligned} \pi_\psi(\psi|\mathbf{x}) &= \pi_\theta\left(\hat{\theta} + \frac{\psi}{\sqrt{n}} \middle| \mathbf{x}\right) \times \frac{1}{\sqrt{n}} \\ &\propto \pi_\theta\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\right) \exp\left\{n\ell_n\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\right)\right\}. \end{aligned}$$

Now taking Taylor series expansions about  $\psi = 0$  gives

$$\begin{aligned} \pi_\theta\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\right) &= \pi_\theta(\hat{\theta}) + \pi'_\theta(\hat{\theta})\frac{\psi}{\sqrt{n}} + O\left(\frac{\psi^2}{n}\right) \\ &= \pi_\theta(\hat{\theta}) \left\{1 + O\left(\frac{\psi}{\sqrt{n}}\right)\right\} \\ n\ell_n\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\right) &= n\left\{\ell_n(\hat{\theta}) + \ell'_n(\hat{\theta})\frac{\psi}{\sqrt{n}} + \frac{1}{2}\ell''_n(\hat{\theta})\frac{\psi^2}{n} + O\left(\frac{\psi^3}{n^{3/2}}\right)\right\} \\ &= n\ell_n(\hat{\theta}) + \frac{1}{2}\ell''_n(\hat{\theta})\psi^2 + O\left(\frac{\psi^3}{\sqrt{n}}\right) \end{aligned}$$

since  $\ell'_n(\hat{\theta}) = 0$  by definition of the maximum likelihood estimate. Therefore, retaining only terms in  $\psi$ , we have

$$\begin{aligned} \pi_\psi(\psi|\mathbf{x}) &\propto \pi_\theta(\hat{\theta}) \left\{1 + O\left(\frac{\psi}{\sqrt{n}}\right)\right\} \exp\left\{n\ell_n(\hat{\theta}) + \frac{1}{2}\ell''_n(\hat{\theta})\psi^2 + O\left(\frac{\psi^3}{\sqrt{n}}\right)\right\} \\ &\propto \exp\{n\ell_n(\hat{\theta})\} \exp\left\{\frac{1}{2}\ell''_n(\hat{\theta})\psi^2\right\} \exp\left\{O\left(\frac{\psi^3}{\sqrt{n}}\right)\right\} \left\{1 + O\left(\frac{\psi}{\sqrt{n}}\right)\right\} \\ &\propto \exp\left\{\frac{1}{2}\ell''_n(\hat{\theta})\psi^2\right\} \left\{1 + O\left(\frac{\psi^3}{\sqrt{n}}\right)\right\} \left\{1 + O\left(\frac{\psi}{\sqrt{n}}\right)\right\} \\ &\propto \exp\left\{\frac{1}{2}\ell''_n(\hat{\theta})\psi^2\right\} \left\{1 + O\left(\frac{\psi}{\sqrt{n}}\right)\right\} \\ &\propto \exp\left\{-\frac{\psi^2}{2[-\ell''_n(\hat{\theta})]^{-1}}\right\} \left\{1 + O\left(\frac{\psi}{\sqrt{n}}\right)\right\}. \end{aligned}$$

Hence

$$\pi_\psi(\psi|\mathbf{x}) \rightarrow k \exp \left\{ -\frac{\psi^2}{2[-\ell_n''(\hat{\theta})]^{-1}} \right\} \quad \text{as } n \rightarrow \infty.$$

Thus, the limiting form of the posterior density for  $\psi$  is that of a  $N(0, [-\ell_n''(\hat{\theta})]^{-1})$  distribution. Hence

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

or, equivalently, since  $n\ell_n(\theta) = \log f(\mathbf{x}|\theta)$

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

as required.

## Comments

1. This asymptotic result can give us a useful approximation to the posterior distribution for  $\theta$  when  $n$  is large:

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

2. The observed information is similar to Fisher's information (3.8). In fact, Fisher's information is the expected value of the observed information, where the expectation is taken over the distribution of  $\mathbf{X}|\theta$ , that is,  $I(\theta) = E_{\mathbf{X}|\theta}[J(\theta)]$ .
3. This limiting result is similar to one for the maximum likelihood estimator in Frequentist Statistics:

$$\sqrt{I(\theta)} (\hat{\theta} - \theta)|\mathbf{x} \xrightarrow{\mathcal{D}} N(0, 1). \quad \text{as } n \rightarrow \infty, \quad (3.10)$$

Note that (3.10) is a statement about the distribution of  $\hat{\theta}$  for fixed (unknown)  $\theta$ , whereas the Theorem is a statement about the distribution of  $\theta$  for fixed (known)  $\hat{\theta}$ .

**Example 3.14**

Suppose we have a random sample from a distribution with probability density function

$$f(x|\theta) = \frac{2x e^{-x^2/\theta}}{\theta}, \quad x > 0, \theta > 0.$$

Determine the asymptotic posterior distribution for  $\theta$ . Note that from Example 3.11 we have

$$\begin{aligned} \frac{\partial}{\partial \theta} \log f(\mathbf{x}|\theta) &= -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n x_i^2, \\ J(\theta) &= -\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{x}|\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n x_i^2 = \frac{n}{\theta^3} \left( -\theta + \frac{2}{n} \sum_{i=1}^n x_i^2 \right). \end{aligned}$$



*...Solution to Example 3.14...*



**Example 3.15**

Suppose we have a random sample from an exponential distribution, that is,  $X_i|\theta \sim \text{Exp}(\theta)$ ,  $i = 1, 2, \dots, n$  (independent). Determine the asymptotic posterior distribution for  $\theta$ . Note that from Example 3.12 we have

$$\begin{aligned}\frac{\partial}{\partial \theta} \log f(\mathbf{x}|\theta) &= \frac{n}{\theta} - n\bar{x}, \\ J(\theta) &= -\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{x}|\theta) = \frac{n}{\theta^2}.\end{aligned}$$




...Solution to Example 3.15...

Recall that, assuming a vague prior distribution, the posterior distribution is a  $Ga(n, n\bar{x})$  distribution, with mean  $1/\bar{x}$  and variance  $1/(n\bar{x}^2)$ . The Central Limit Theorem tells us that, for large  $n$ , the gamma distribution tends to a normal distribution, matched, of course, for mean and variance. Therefore, we have shown that, for large  $n$ , the asymptotic posterior distribution is the same as the posterior distribution under vague prior knowledge. Not a surprising result!

**Example 3.16**

Suppose we have a random sample from a  $N(\mu, 1/\tau)$  distribution (with  $\tau$  known). Determine the asymptotic posterior distribution for  $\mu$ . Note that from Example 3.13 we have

$$\begin{aligned}\frac{\partial}{\partial \mu} \log f(\mathbf{x}|\mu) &= n\tau(\bar{x} - \mu), \\ J(\mu) &= -\frac{\partial^2}{\partial \mu^2} \log f(\mathbf{x}|\mu) = n\tau.\end{aligned}$$

 ...Solution to Example 3.16...

Again, we have shown that the asymptotic posterior distribution is the same as the posterior distribution under vague prior knowledge.