MAS2903

Introduction to Bayesian Methods

Semester 2, 2018—2019

Lecturer: Dr. Lee Fawcett

School of Mathematics, Statistics & Physics

MAS2903: Introduction to Bayesian Methods

Lecturer: Dr. Lee Fawcett

Office: Room 2.07, Herschel Building Phone: 208 7228

Email: lee.fawcett@ncl.ac.uk

Classes

- There will be two *lectures* each week: Mondays at 11am and Tuesdays at 3pm, both in Lecture Theatre 1 of the Herschel Building.
- In the first two weeks there will be a third lecture on Thursdays at 2pm, also in Lecture Theatre 1 of the Herschel Building.
- From the third week onwards *problems classes* and *drop-in sessions* take place in alternate weeks on Thursdays at 2pm in Lecture Theatre 1 of the Herschel Building.
- *Computer practicals* will take place in teaching weeks 4 (before Easter) and 10 (after Easter). These will be on Fridays at 1pm in the Herschel PC cluster. A reminder will be given the week before!
- *Office hours:* I will be available to see MAS2903 students on Mondays and Thursdays 12-1pm (my office).

What?	When?		Where?	How often?
Lecture	Monday	11–12	Herschel: LT1	Every week
	Tuesday	3–4	Herschel: LT1	Every week
Problems class/ Drop–in session	Thursday	2–3	Herschel: LT1	Every week has one or the other
Computer practical	Friday	1–2	Herschel: PC cluster	Some even weeks
Office hours	Monday	12–1	My office	Every week
	Thursday	12–1		

Assessment

- **85%** exam in May/June
- **15%** in course assessment:
  - Mid–semester test (10%): **Friday 15th March, 1pm** (split between several rooms – details to follow by email)
  - 2 assignments (5% in total), each one having a mixture of written questions and practical work. Deadlines – **Fridays** at **4pm** on the following dates:

    8th March, **3rd May**.

Recommended textbooks

- "*Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*" - James Stone
- "*Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*" - John Krushke
- "*Bayesian Statistics: An Introduction*" - Peter Lee

"Bayes' rule" is a good introduction to the main concepts in Bayesian statistics but doesn't cover everything in this course. The other books are broader references which go well beyond the contents of this course.

Other stuff

- Notes (with gaps) will be handed out in lectures – you should fill in the gaps during lectures.
- A summarised version of the notes will be used in lectures as slides.
- The notes are available on Blackboard. The slides will be posted after each topic is finished, along with any other course material – such as model solutions to assignment questions, supplementary handouts etc.
- There is also a course webpage (just in case Blackboard breaks):

  www.mas.ncl.ac.uk/∼nlf8/teaching/mas2903

- Please check your University email account regularly, as course announcements will often be made via email.

Before starting the course 'proper', we will spend the first lecture going over some background to the subject of Bayesian Statistics. Although the material in the rest of this handout is non–examinable, it will help to 'set the scene' for the course as well as motivate the techniques we will develop. Let's make a start...

Preface to lecture notes: What is Bayesian Statistics?

There are two main approaches to statistics: *frequentist* (or classical) statistics and *Bayesian* statistics. All of the statistics teaching you've encountered so far is likely to be about frequentist methods. Bayesian methods are substantially different and can feel quite strange to start with. So, before starting the main course material, this *non-examinable* preface explains the concepts behind Bayesian statistics and how they differ from frequentist approaches.

One way of defining statistics is as a way to learn about the world from some data which is subject to random variation. For example, in *climate science* we want to learn about the climate given some imperfect measurements. Some statistical questions that we might ask in this field are:

- What is our best estimate of the world temperature this year?
  This is referred to as a *point estimation* problem.
- What is a plausible range of values?
  This is referred to as an *interval estimation* (or uncertainty quantification) problem.
- What will the temperature be in 100 years?
  This is a *prediction* problem.
- Is the climate warming?
  This is a *hypothesis testing* problem.

Frequentist and Bayesian methods address all these types of problem: point estimation, interval estimation, prediction and hypothesis testing. But they use different approaches to do so. Some typical frequentist approaches include:

- Least squares (point estimation)
- Maximum likelihood (point estimation)
- Confidence intervals (interval estimation)
- Test statistics and p-values (hypothesis testing)

Bayesian statistics doesn't use any of these familiar methods! However, there are some connections between them and their Bayesian alternatives which will be explored later in the course. For example we'll see that the idea of the *likelihood function* is very important in Bayesian statistics.

Motivating example

Air France Flight 447 disappeared over the Atlantic Ocean on June 1st 2009, during an overnight flight from Rio de Janeiro to Paris. A search and rescue mission was launched the following morning. After five days some floating wreckage was found, but there was no sign of the "black box" recorders which would reveal what had happened to the flight, and the search stalled.

The black boxes could be anywhere within an area of the South Atlantic the size of Switzerland (6,500 square miles). The "mid-Atlantic ridge" on the ocean floor lies between two tectonic plates and is just as mountainous as Switzerland! It's also so remote that scientists have not yet charted the sea-bed. The search method used was sonar-detectors, emitting sound waves which would bounce back once they hit something. However, after two years of meticulously searching an area north of the planes trajectory (after analysing debris drift), nothing had been found.

Metron, Inc., of Retson, Virginia had been performing Bayesian analyses to help the search effort. Included in their analysis were:

1. Data from 9 previous airline crashes involving loss of pilot control - reduced search area to 1,600 square miles.

2. Expert opinions on the credibility of the flight data.

3. Expert opinions about whether or not the black box 'pingers' might have become damaged on impact.

4. Positions/recovery times of bodies found drifting - expert opinions assigned to the reliability of this data because of the turbulent equatorial waters.

5. Expert information from oceanographers: Sea state, visibility, underwater geography,...

All were combined through Bayes Theorem to give a probability map of the most likely locations to search. In April 2011 they decided to try assuming that the 'pingers' in item 3 were probably damaged. One week later the black boxes were found![1]

[1] For more details see `https://www.technologyreview.com/s/527506/how-statisticians-found-air-france-flight-447-two-years-after-it-crashed-into-atlantic/`.

The frequentist/Bayesian debate

This example illustrates some key features of the Bayesian approach:

- It incorporates expert opinion.
- It outputs a probability distribution over the possible choices.
- The calculations required are all based on Bayes Theorem.

In contrast in frequentist statistics it is difficult (but not impossible) to incorporate opinions, and the output is not in the form of a probability distribution. For example a 95% frequentist confidence interval is a range which will contain the true value 95% of the time if the analysis is repeated a large number of times for different data obtained under the same process[2]. In contrast a Bayesian 95% interval estimate (see Chapter 4) is a range with a 95% probability of containing the true value. This Bayesian version is arguably much easier to interpret.

There has been a decades-long argument about whether Bayesian or frequentist methods are better. Both sides have some good points. For example, there is the argument we've just made about Bayesian interval estimates being easy to interpret. On the other hand one of the main arguments against the Bayesian approach is that it is *subjective*. This is because its results are based on expert opinions, and if we got a different opinion from another expert our results would change even though the data are the same! Bayesians have counter-arguments. For example many would argue that subjectivity is in fact a desirable feature, or that frequentist methods also have hidden subjective elements e.g. using 5% as a significance threshold rather than some other value. The philosophical debate seems unlikely to end any time soon!

In practice many problems are better suited to be tackled with methods from one approach or the other. For example Bayesian statistics was well suited to the Air France example as incorporating expert knowledge was particularly important in the solution. Therefore it's good to be familiar with both Bayesian and frequentist methods.

The history of Bayesian statistics

The reason why we use the word "Bayesian" is because Bayes' Theorem is crucial for statistical analysis if we adopt the Bayesian approach. Thomas Bayes (1702–1761; see picture below) was a Presbyterian minister in Tunbridge Wells, Kent. Bayes' solution to a problem in probability theory was presented in the *Essay towards Solving a Problem in the Doctrine of Chances*, published posthumously by his friend Richard Price in the *Philosophical Transactions of the Royal Society of London*. This paper gave us Bayes' Theorem. Its modern form is due to Pierre-Simon Laplace who generalised Bayes' work

[2]Here 95% is the *long-run frequency* that the interval contains the correct value. This kind of property is where the name "frequentist" comes from.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

Figure 1: Portrait believed to be of Thomas Bayes, and letter from Richard Price to the Philosophical Transactions of the Royal Society of London

and produced the first theory of Bayesian inference, which was referred to as "inverse probability". Laplace's theory had weaknesses[3] and in the early 20th century it was replaced by the more rigorous frequentist approach. R. A. Fisher was one of the leading creators of this approach. He was also the first to use the term "Bayesian statistics".

Nonetheless there was still some interest in Bayesian methods. For example, Alan Turing's group invented some new Bayesian methods to help break the Enigma code in World War II. Before and, increasingly, after the war a few researchers worked on developing a rigorous theory of Bayesian statistics.

Bruno de Finetti (1906–1985) was an Italian probabilist who developed his ideas on *subjective probability*[4] from the 1920s onwards, drawing upon ideas from H. Jeffreys, I.J. Good (one of the Enigma codebreakers) and B.O. Koopman. His classic book on the topic is *The Theory of Probability* (1974).

Dennis Lindley (1923–2013) was a leading British Bayesian. In his early career, he worked to find a mathematical basis for the subject of statistics. In 1954, Lindley met Leonard Savage and both found a deeper justification for statistics in Bayesian theory, turning into critics of the classical statistical inference they had hoped to justify.

Many of these theoretical advances are covered in this course. However, Bayesian statistics was still little used in practice. The difficulty was that applying Bayes theorem often

[3]For example it relied mainly on *flat priors*, which will be discussed and criticised in Chapter 3.
[4]Covered in Chapter 1.

Figure 2: Pierre-Simon Laplace (left) and R. A. Fisher (right)

became very mathematically challenging, resulting in problems that were too difficult or impossible to solve by hand.

Recent developments

In the 1990s a solution to these mathematical problems was developed, making use of the emergence of more powerful computers and "Markov chain Monte Carlo" (MCMC) numerical algorithms. If you take 3rd and 4th year Bayesian courses you will be introduced to these techniques. MCMC, and subsequent developments, have revolutionised the use of Bayesian Statistics, to the extent that today Bayesian data analyses are as popular as frequentist approaches and are routinely used in fields as diverse as artificial intelligence, biology, astrophysics and sociology.

Newcastle's contribution

Since 1980, the number of academic staff in Mathematics & Statistics at Newcastle publishing advanced research using Bayesian methods has increased dramatically. In the 1980s, there was only one Bayesian at Newcastle. Now there are at least 12. Our Bayesian research topics include environmental extremes, genetics, and modelling biological systems at a molecular level.

Chapter 1

Introduction

Consider the following three experiments.

Experiment 1: Fisher's 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

$$\text{win £}c \text{ if } A \text{ occurs and £0 if } A^{\mathrm{c}} \text{ occurs.}$$

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 £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 £1 on the spin of a coin to win £2 if it lands heads may refuse to to bet if the stakes are raised to £1000 – most people are *risk–averse*. Therefore, we shall restrict our attention to the $c = 1$ case: pay £p for the bet

$$\text{win £1 if } A \text{ occurs and £0 if } A^{\mathrm{c}} \text{ occurs.}$$

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 £0.50. However, if you weren't honest you might try to buy the bet for any amount less than £0.50, say £0.20. If you were hosting the bet, you would take the bet for any amount more than £p, say for £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(Newcastle Utd win the Championship this season) could only be determined using a subjective assessment.
2. The probability Pr(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(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

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

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

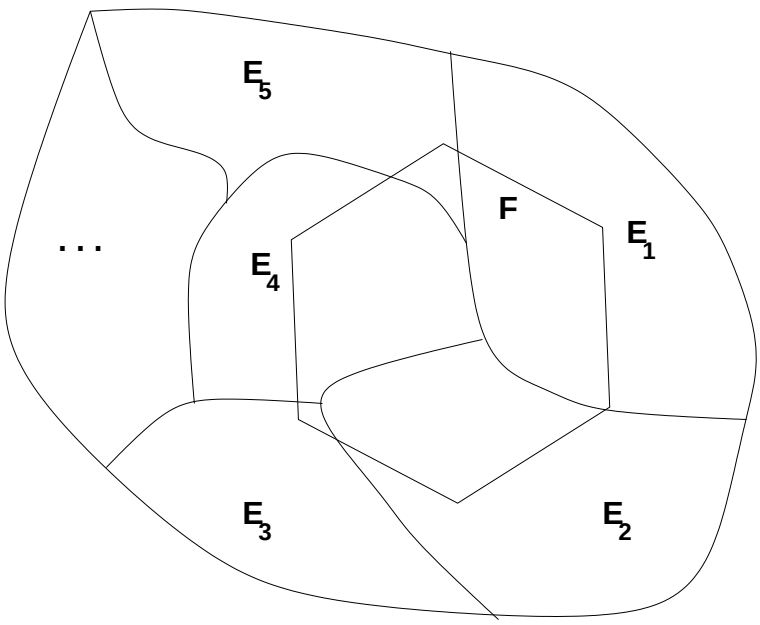

Figure 1.1: Diagram of a partition $E_1, E_2, \ldots, E_n$ and an event F

Law of Total Probability

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

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

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.

(a) Calculate the probability that a person who tests positive actually has the disease.

(b) 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

$$x_1 : \text{overheating only},$$
$$x_2 : \text{irregular traction only},$$
$$x_3 : \text{both symptoms}.$$

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 x_1	I/T x_2	Both 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ and we decide to model the data using a probability (density) function $f(\boldsymbol{x}|\theta)$. This p(d)f describes how likely different data $\boldsymbol{x}$ are to occur given a value of the (unknown) parameter θ. However, once we have observed the data, $f(\boldsymbol{x}|\theta)$ tells us how likely different values of the parameters θ are: it is then known as the *likelihood function* for θ. In other courses you may have seen it written as $L(\theta|\boldsymbol{x})$ or $L(\theta)$ but, whatever the notation used for the likelihood function, it is simply the joint probability (density) function of the data, $f(\boldsymbol{x}|\theta)$, regarded as a function of θ rather than of $\boldsymbol{x}$.

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

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta), \tag{1.1}$$

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

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f_{X}(x_i|\theta). \tag{1.2}$$

In this course, we will not 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 (1.2) form of the likelihood function.

Example 1.5

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

✎...*Solution to Example 1.5*...

Example 1.6

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

✎...*Solution to Example 1.6*...

Example 1.7

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a normal distribution: $X_i|\mu, \sigma \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ (independent). Determine the likelihood function for (μ, σ).

✎ *...Solution to Example 1.7...*

1.4 Sufficiency

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

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

Notice that this likelihood function depends on the data only through $\prod_{i=1}^n (x_i!)^{-1}$ and $\bar{x}$. Further, in $f(\boldsymbol{x}|\theta)$, θ only "interacts" with $\bar{x}$ — the other term simply scales $f(\boldsymbol{x}|\theta)$ — so that, for example, the point at which $f(\boldsymbol{x}|\theta)$ 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 $T(\boldsymbol{X})$ is *sufficient* for θ if $f(\boldsymbol{x}|T(\boldsymbol{X}) = t)$ 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 $X_i|\theta \sim Po(\theta)$, $i = 1, 2$ (independent). Show that $T = X_1 + X_2$ is sufficient for θ. Note that $T|\theta \sim Po(2\theta)$.

✎ *...Solution to Example 1.8...*

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

Definition 1.5

The statistics $\boldsymbol{T}(\boldsymbol{X}) = \big(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_k(\boldsymbol{X})\big)^T$ are (jointly) *sufficient* for θ if $f(\boldsymbol{x}|\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t})$ does not depend on θ.

Factorisation Theorem

Under certain regularity conditions

$$\boldsymbol{T}(\boldsymbol{X}) \text{ is sufficient for } \theta \iff f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\, g(\boldsymbol{t}(\boldsymbol{x}), \theta)$$

for some functions h and g.

Example 1.9

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

...Solution to Example 1.9...

✎ *...Solution to Example 1.9...*

Example 1.10

Suppose we have a random sample from an exponential distribution: $X_i|\theta \sim Exp(\theta),\ i = 1, 2, \ldots, n$ (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: $X_i|\mu, \sigma \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ (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 1.4, where we had a random sample from an exponential distribution, we found that $T = \Sigma X_i$ was sufficient for θ. However, (obviously) the whole data $\boldsymbol{X}$ is sufficient for θ since

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= 1 \times f(\boldsymbol{x}|\theta) \\ &= h(\boldsymbol{x})\, g(\boldsymbol{x}, \theta), \end{aligned}$$

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

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= 1 \times \theta^n \exp(-n\theta\bar{x}) \\ &= h(\boldsymbol{x})\, g_1\left(\bar{x}, \theta\right), \text{ or} \\ &= h(\boldsymbol{x})\, g_2\left((\bar{x})^2, \theta\right), \text{ or} \\ &= h(\boldsymbol{x})\, g_3\left((\bar{x})^3, \theta\right), \text{ or} \\ &= h(\boldsymbol{x})\, g_4\left((\bar{x})^4, \theta\right), \text{ or} \\ &= \cdots \end{aligned}$$

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

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

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= 1 \times \theta^n \exp\left(-n\theta\bar{x}\right) \\ &= h(\boldsymbol{x})\, g_1(\bar{x}, x_1, \theta), \text{ or} \\ &= h(\boldsymbol{x})\, g_2(\bar{x}, x_2 \sin x_6, \theta), \text{ or} \\ &= \cdots \end{aligned}$$

where $h(\boldsymbol{x}) = 1$ and $g_j(t, \boldsymbol{s}, \theta) = \theta^n \exp\left(-\theta t\right)$.

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

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

Bayes Theorem

The posterior probability (density) function for θ is

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

where

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

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

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

Thus, to obtain the posterior distribution, we need:

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

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

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

Definition 2.1 (Continuous Uniform distribution)

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

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

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

You should remember from MAS1604 that

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

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

Definition 2.2 (Beta distribution)

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

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

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

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

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

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

where

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

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

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

For example,

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

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

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

Also

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

Definition 2.3 (Gamma distribution)

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

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

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

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

Also

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

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

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

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

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

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

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

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

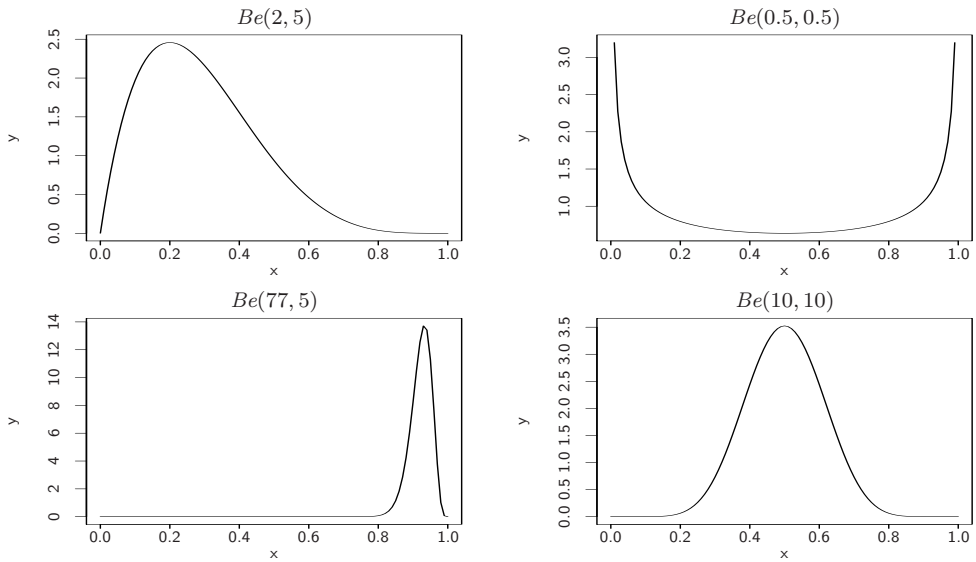

Figure 2.1: Plots of $Be(a,b)$ densities for various values of (a,b).

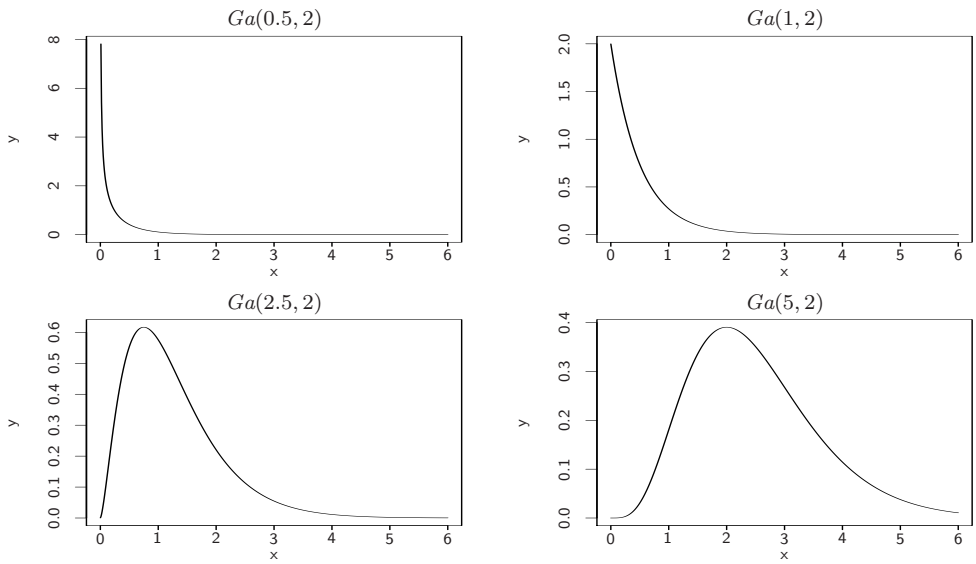

Figure 2.2: Plots of $Ga(a, b = 2)$ densities, for various values of a.

2.2 Bayes' Theorem for distributions in action

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

Example 2.1

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

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

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

✎ *...Solution to Example 2.1...*

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

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

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

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

Summary:

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

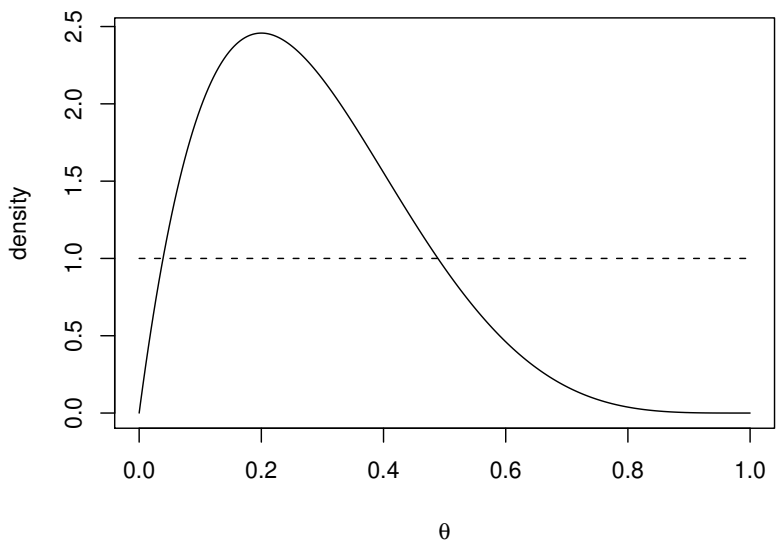

Figure 2.3: Prior (dashed) and posterior (solid) densities for $\theta = \Pr(\text{Head})$

Example 2.2

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

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

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

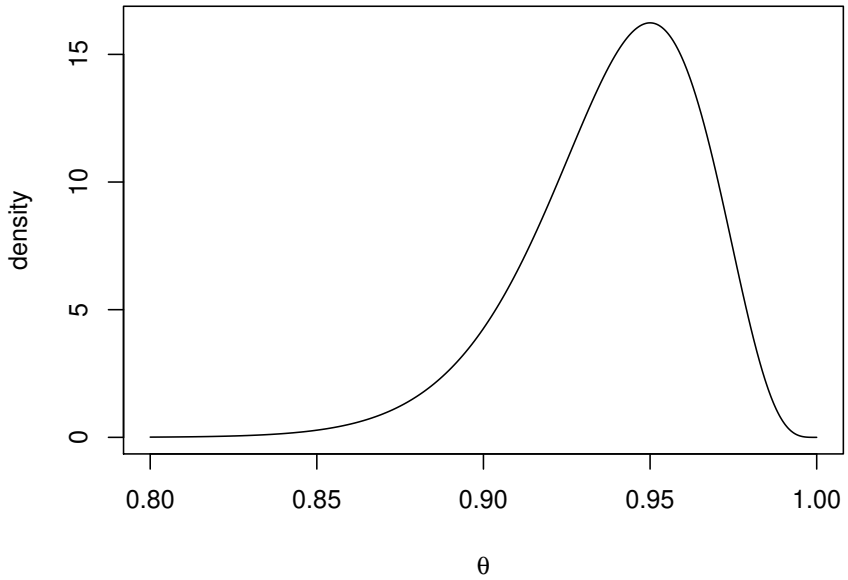

Figure 2.4: Prior density for the music expert's skill.

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

...Solution to Example 2.2...

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

Summary:

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

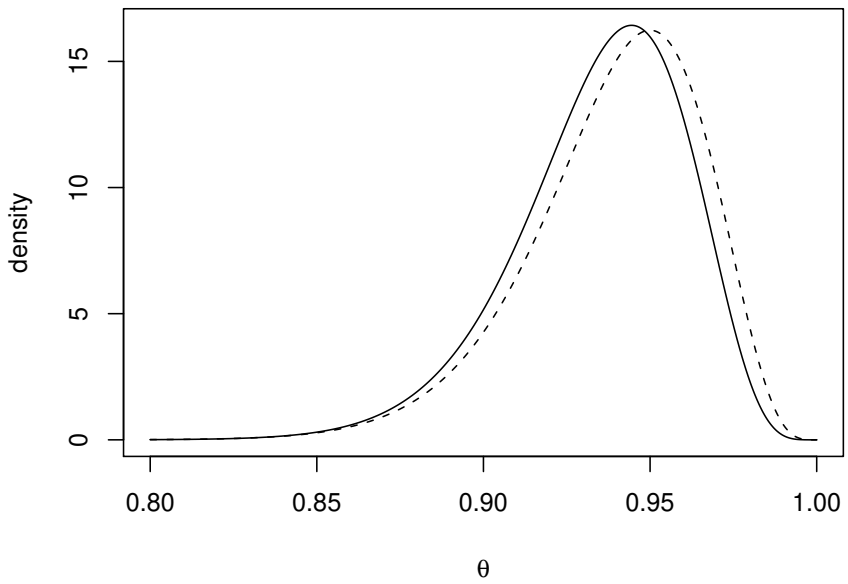

Figure 2.5: Prior (dashed) and posterior (solid) densities for the music expert's skill.

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

Table 2.1: Changes in beliefs about θ.

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

Example 2.3

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

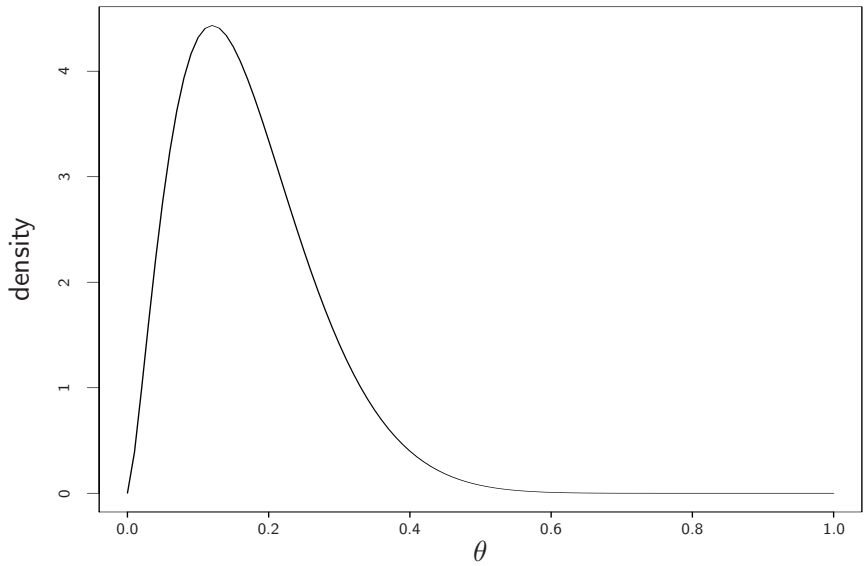

Figure 2.6: Max's prior density.

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

...Solution to Example 2.3...

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

Summary:

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

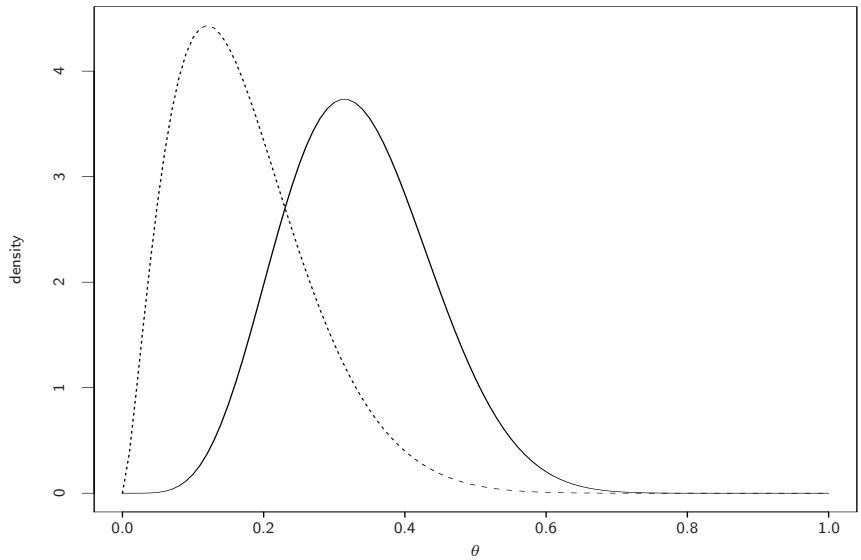

Figure 2.7: Prior (dashed) and posterior (solid) densities for Max's problem.

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

Table 2.2: Changes in beliefs about θ.

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

Example 2.4

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

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

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

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

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

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

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

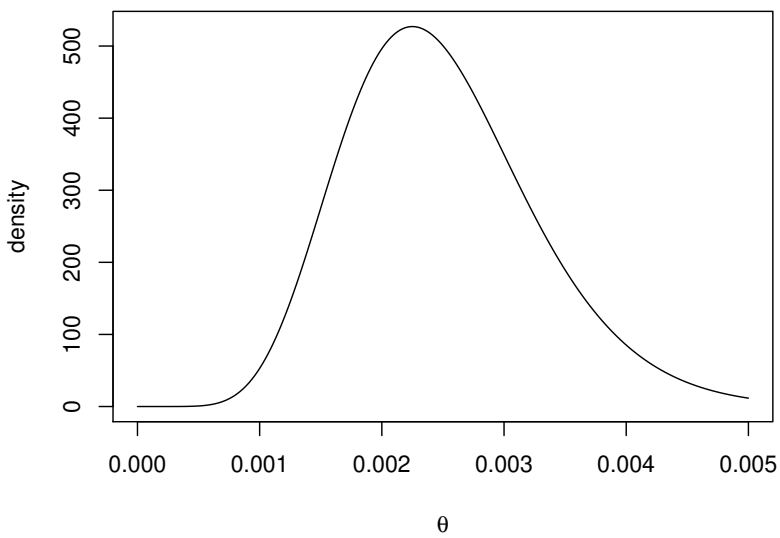

Figure 2.8: Prior density for the earthquake rate θ

✎ *...Solution to Example 2.4...*

Summary:

The data have updated our beliefs about θ from a $Ga(10, 4000)$ distribution to a $Ga(30, 13633)$ distribution. Plots of these distributions are given in Figure 2.9, and Table 2.4 gives a summary of the main changes induced by incorporating the data. Notice that, as the mode of the likelihood function is close to that of the prior distribution, the information in the data is consistent with that in the prior distribution. This results in a reduction in variability from the prior to the posterior distributions. The similarity between the prior beliefs and the data has reduced the uncertainty we have about the likely earthquake rate θ.

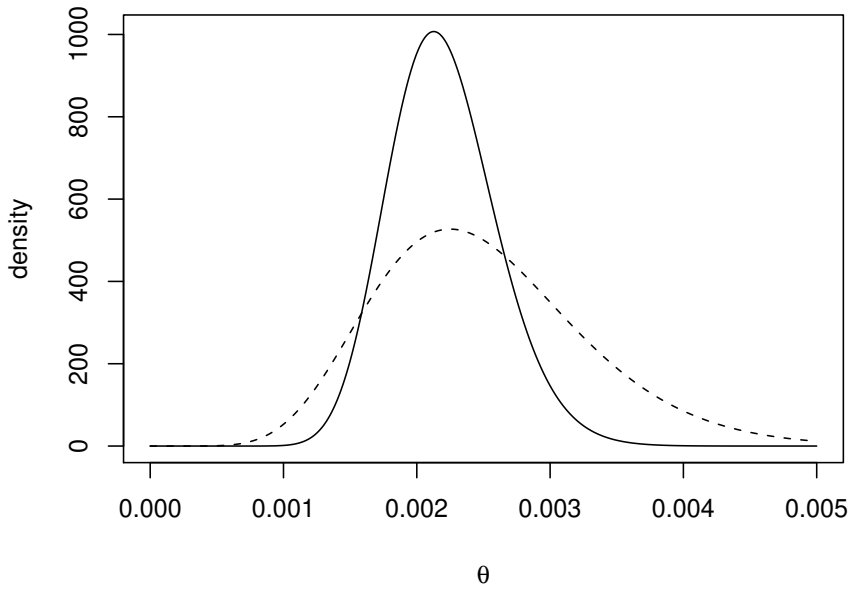

Figure 2.9: Prior (dashed) and posterior (solid) densities for the earthquake rate θ.

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

Table 2.4: Changes in beliefs about θ.

Example 2.5

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

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

Determine the posterior distribution for θ.

...Solution to Example 2.5...

Summary:

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

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

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

Table 2.5: Changes in beliefs about θ.

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

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

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

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

Conjugacy

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

Definition 2.4

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

Notice that the conjugate family depends crucially on the model chosen for the data $\boldsymbol{x}$. So we say (for example) that the Gamma distribution is a conjugate prior for the exponential model.

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

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

Some examples of conjugacy are as follows:

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

Example 2.6

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

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

Determine the posterior distribution for μ.

✎ *...Solution to Example 2.6...*

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

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

Summary:

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

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

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

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

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

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

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

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

The changes in our beliefs about μ are summarised in Table 2.6. Notice that the posterior mean is greater than the prior mean if and only if the likelihood mode (sample mean) is greater than the prior mean, that is

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

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

	Prior (2.19)	Likelihood (2.20)	Posterior (2.22)
$Mode(\mu)$	b	$\bar{x}$	$(db + n\tau\bar{x})/(d + n\tau)$
$E(\mu)$	b	–	$(db + n\tau\bar{x})/(d + n\tau)$
$Precision(\mu)$	d	–	$d + n\tau$

Table 2.6: Changes in beliefs about μ.

Example 2.7

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

✎ *...Solution to Example 2.7...*

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

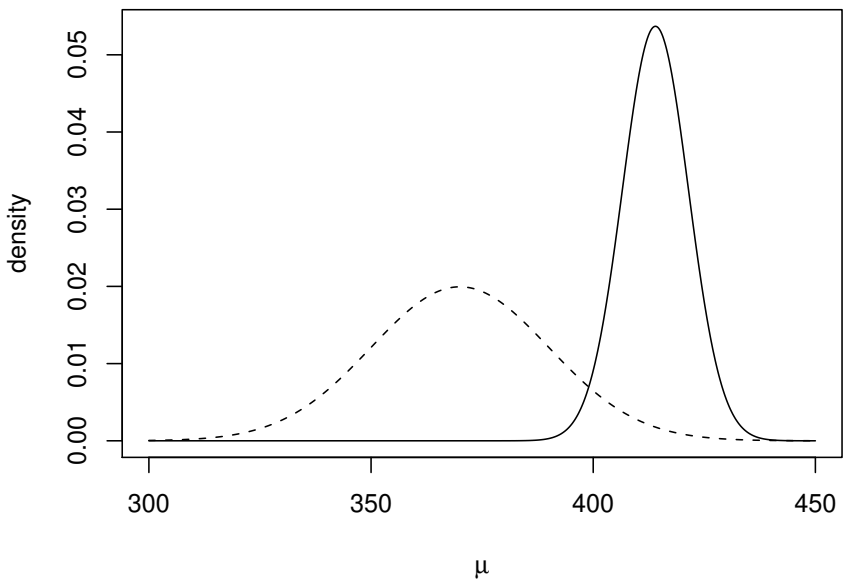

Figure 2.10: Prior (dashed) and posterior (solid) densities for the age of the rock

2.3 Posterior distributions and sufficient statistics

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

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

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

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

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

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

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

Hence

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

However, applying Bayes Theorem to the data t gives

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

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

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

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

Example 2.8

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

✎ *...Solution to Example 2.8...*

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

Example 2.9

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

✎ *...Solution to Example 2.9...*

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 θ. For example, why did we use a $Be(77, 5)$ distribution for θ in the music expert example (page 31)? Why did we use a $Be(2.5, 12)$ distribution for θ in the example about the video game pirate (page 34)? And why did we assume a $Ga(10, 4000)$ distribution for θ in the earthquake example (page 37)? *Prior elicitation* – the process by which we attempt to construct the most suitable prior distribution for θ – 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 permissable range for θ 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 θ. 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 θ (*vague prior knowledge*). Alternatively, a prior which assumes all values of θ are equally likely could be used to represent complete *prior ignornace*, as in the example about the possibly biased coin on page 28.

In this chapter we will also consider the construction of priors for θ 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 exponenetial 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 θ, 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 6.25×10^{-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 θ.

Before conducting the experiment, we were told that the expert is very competent; in fact, we were told that θ 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, θ 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 θ?

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

$$\frac{a-1}{a+b-2} = 0.95 \tag{3.1}$$

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

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

$$\int_0^{0.8} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} d\theta = 0.0001, \qquad \text{i.e.}$$

$$\int_0^{0.8} \frac{\theta^{(19b-18)-1}(1-\theta)^{b-1}}{B(19b-18,b)} d\theta = 0.0001, \tag{3.3}$$

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:

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

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 Elcitation 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 θ 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 θ using the *trial roulette method*. Here, we divide the sample space for θ 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 θ 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, θ 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 α and β 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 θ; twice as many chips in one bin compared to another would mean that the expert believes that θ 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.

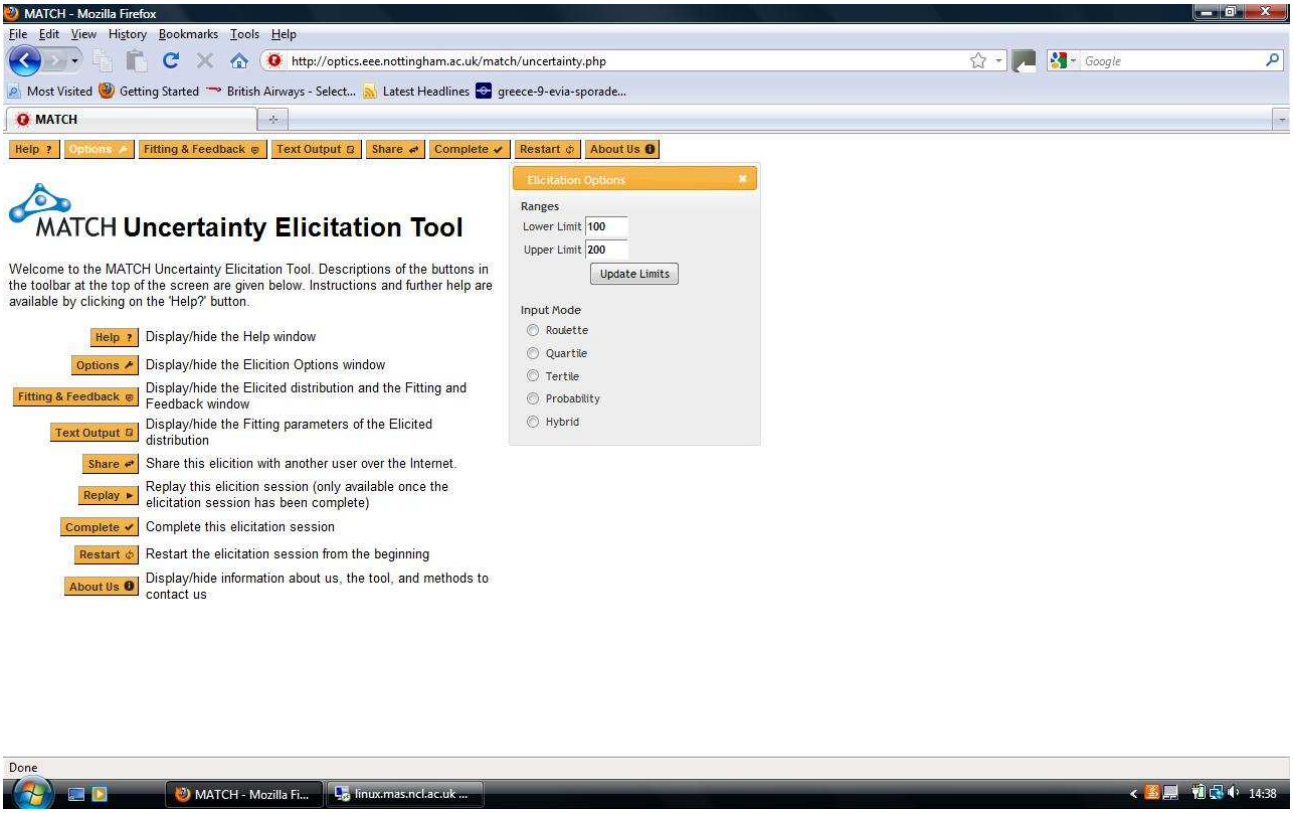

Figure 3.1: The homepage of the *MATCH Uncertainty Elicitation Tool*.

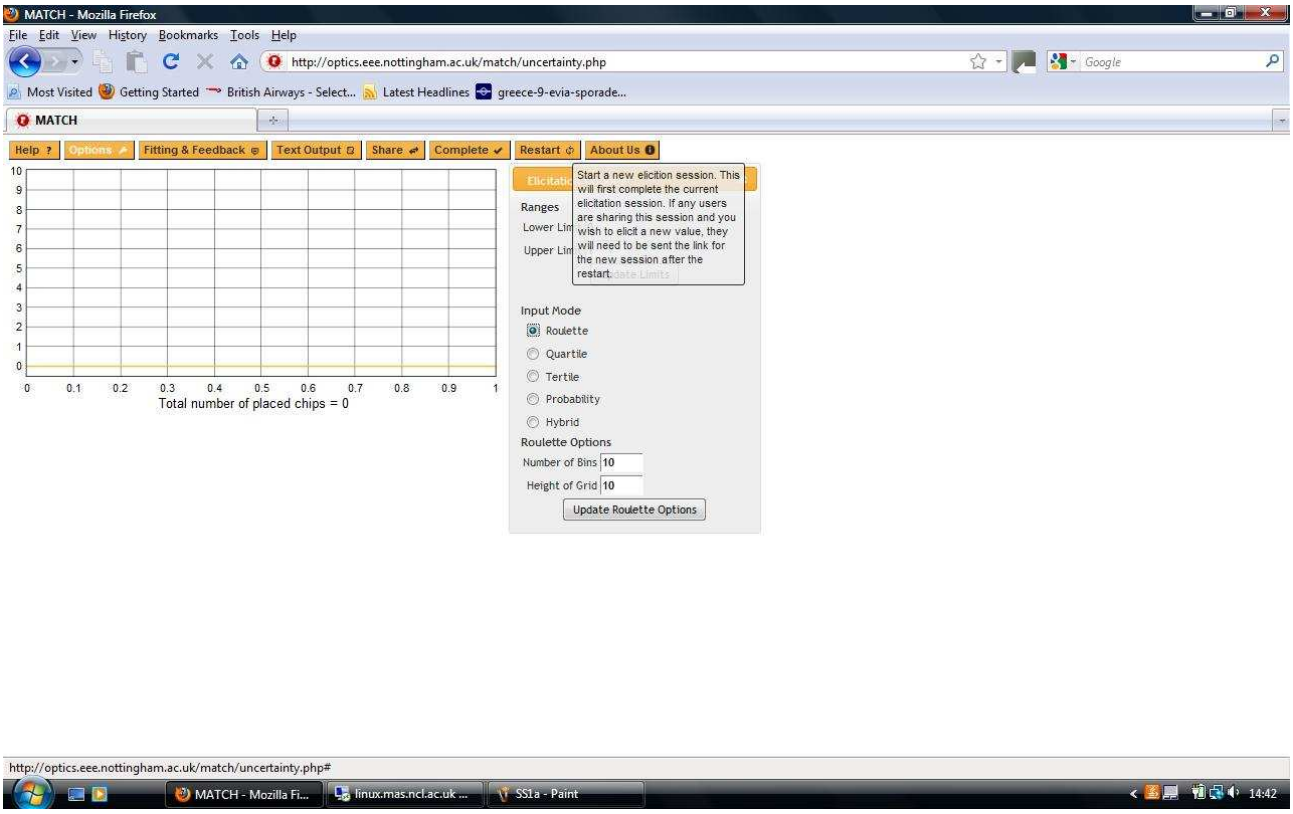

Figure 3.2: The *Roulette* option in *MATCH*.

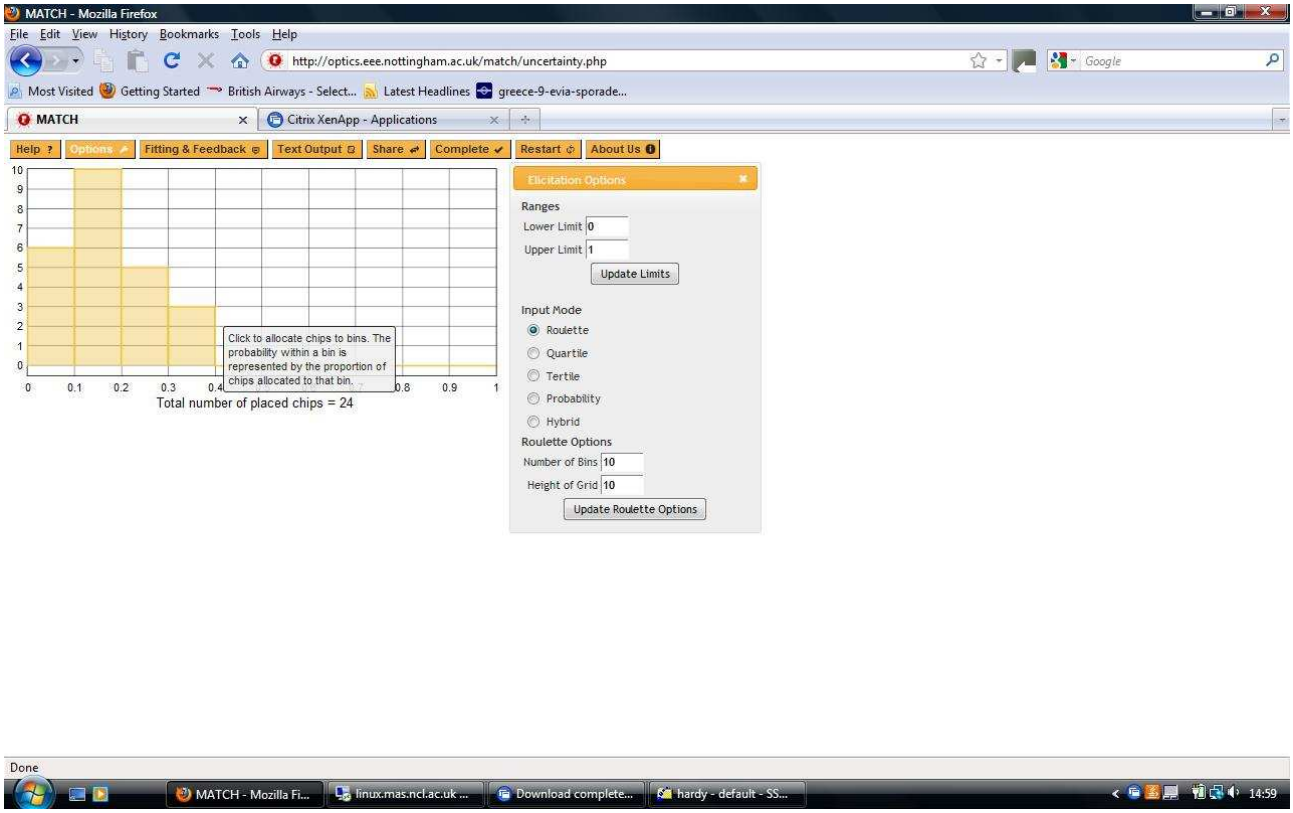

Figure 3.3: Placing chips in bins using the *Roulette* option in *MATCH*.

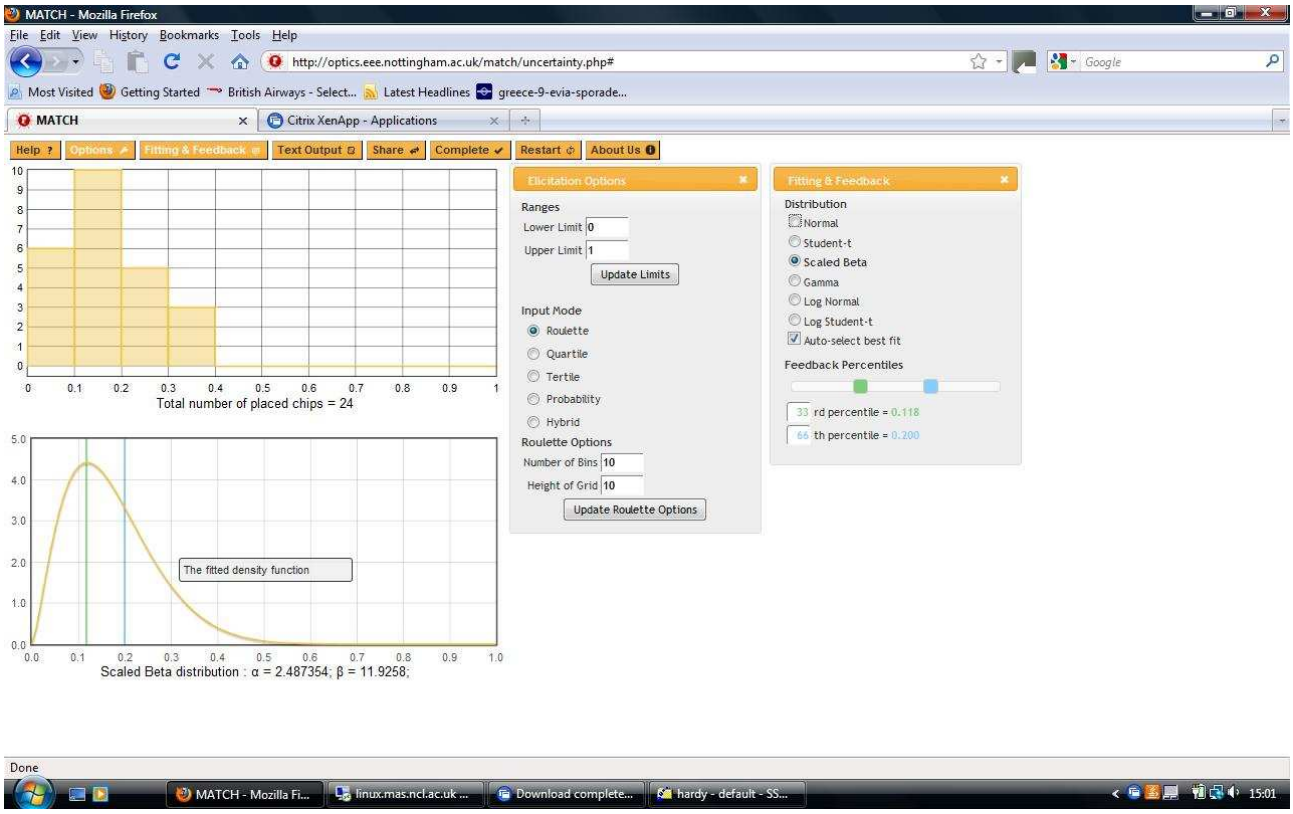

Figure 3.4: Fitting and Feedback in *MATCH*.

Example 3.4 (Bisection method)

The *bisection method* is another method of prior elcitation. This method requires the expert to be able to split the range of allowable values for θ 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 θ; 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 θ; 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, θ (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 elicitng a suitable prior distribution for θ 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 θ, respectively. We now attempt to elicit the median and quartiles for θ 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 θ 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 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.

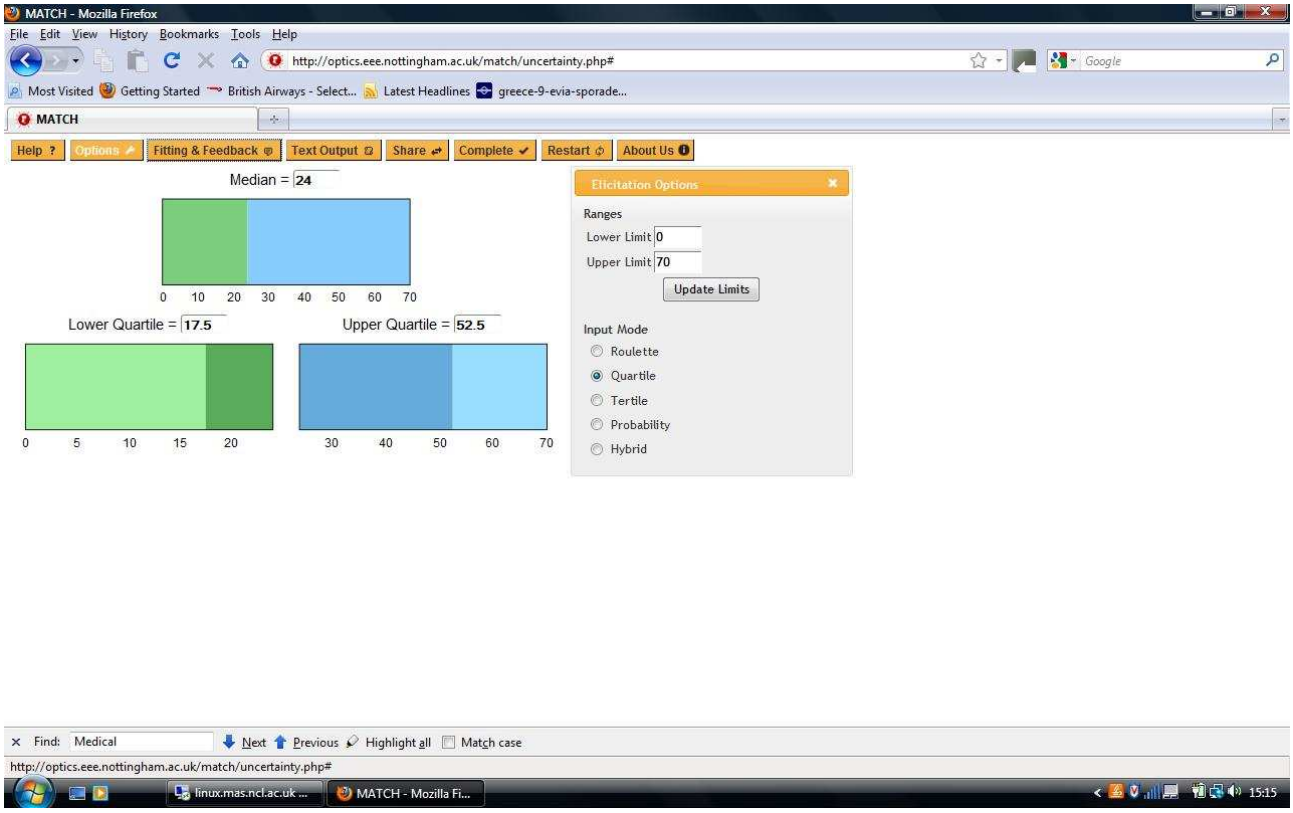

Figure 3.5: Entering the lower and upper limits, and the elicited median, in *MATCH*.

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 northely 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.

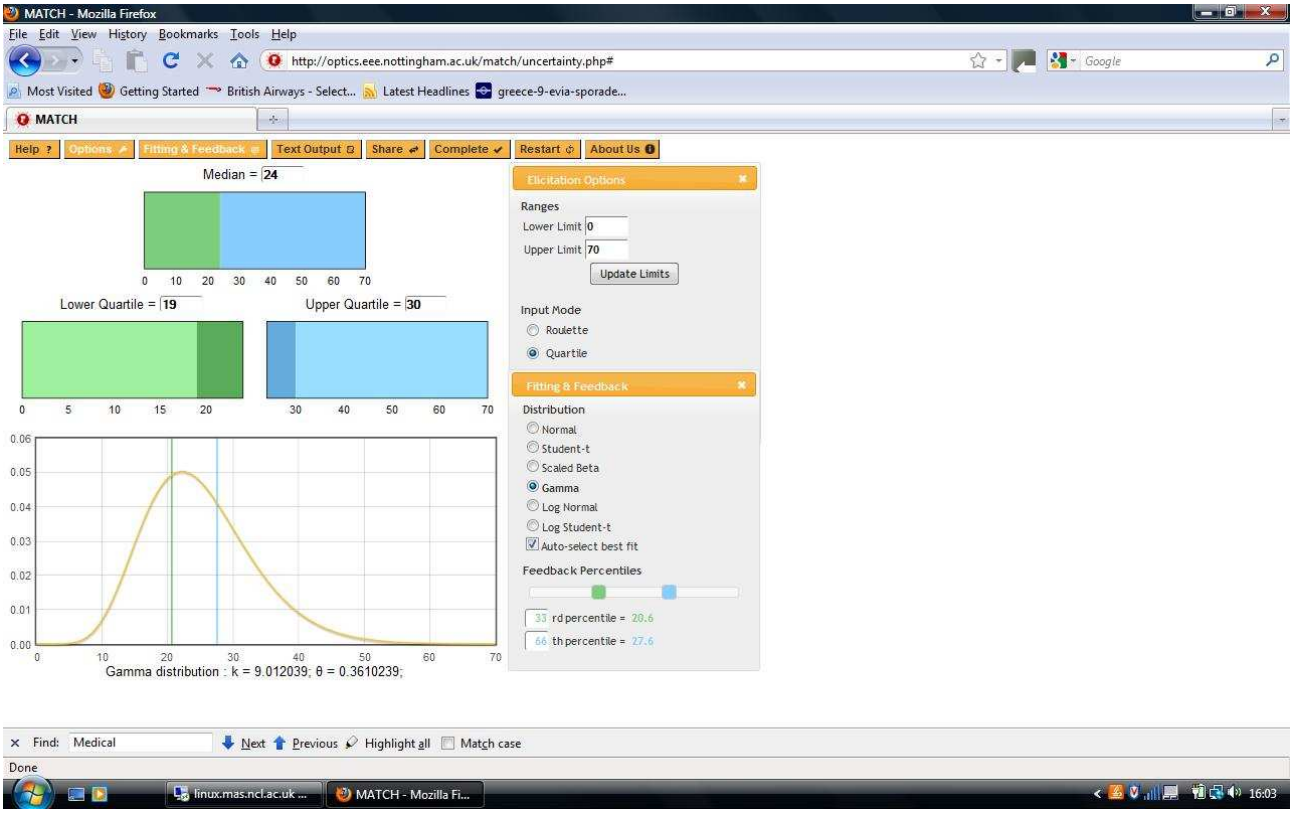

Figure 3.6: Entering the lower and upper quartiles to obtain a fitted distribution for θ in *MATCH*.

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 θ. 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 diminshing 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."

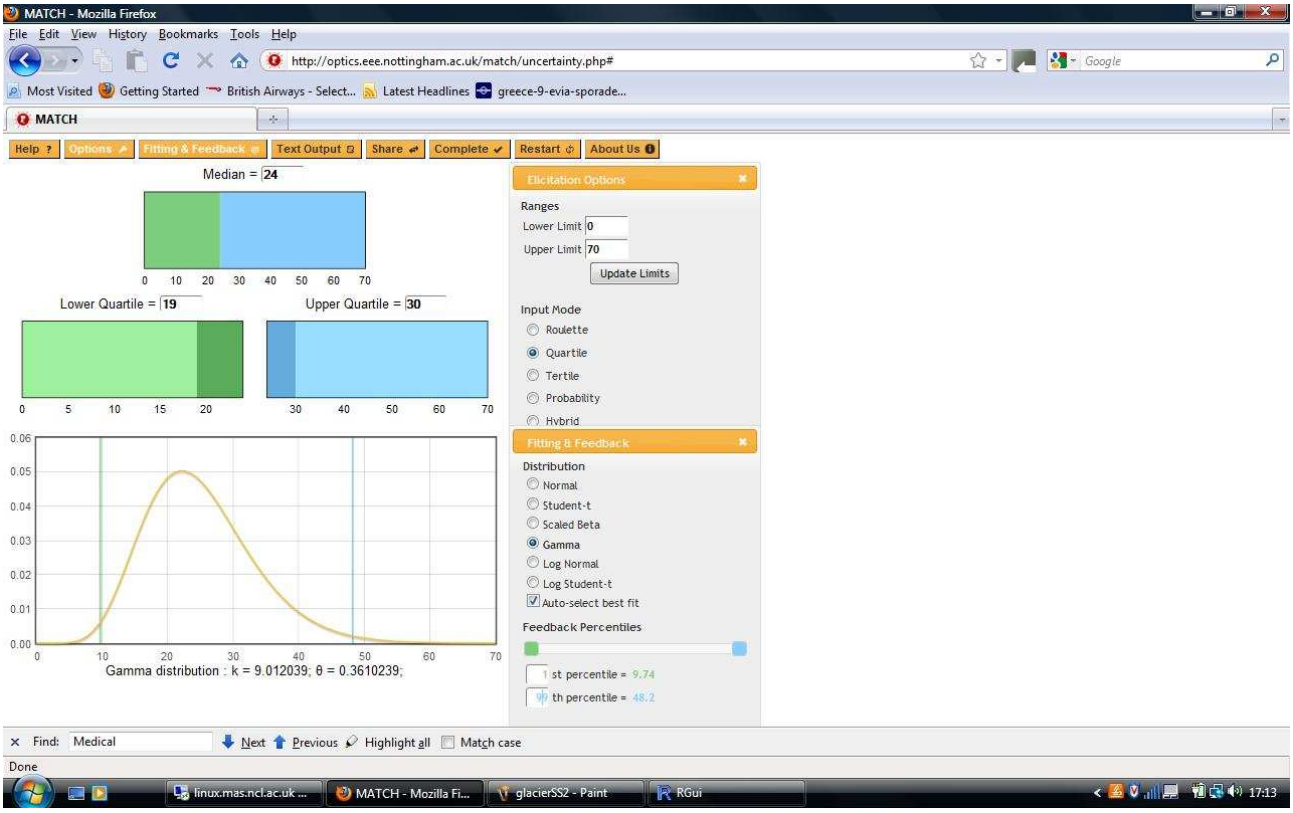

Figure 3.7: Obtaining tail probabilities in *MATCH*: here, we ask for the 1st percentile and the 99th percentile.

Example 3.5

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

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

(a) Obtain the likelihood function for θ given the parameter κ and some observed data $y_1, y_2, \ldots, y_n$ (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 θ.

✎ *...Solution to Example 3.5(b)...*

(c) Using the elicted prior for the rate of retreat we obtained from the expert glaciologist in Example 3.4, and assuming κ 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...*

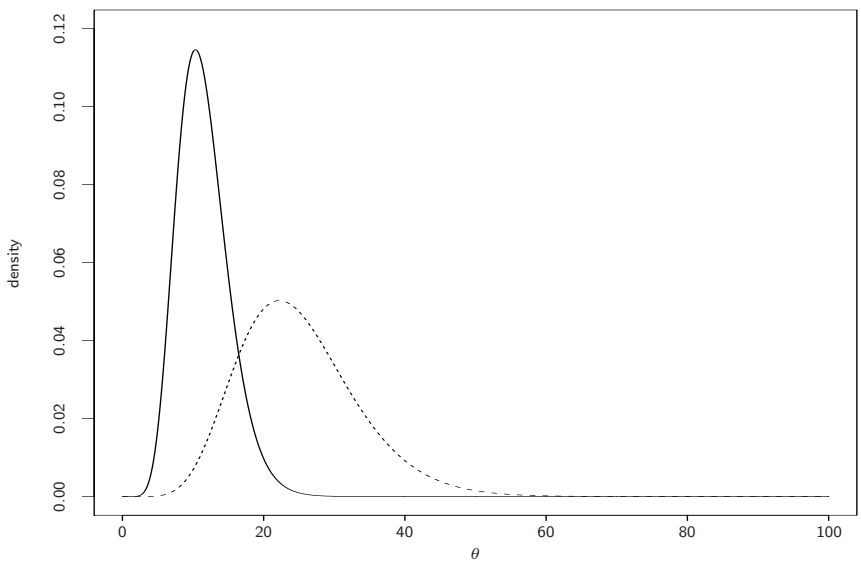

Figure 3.8: Prior (dashed) and posterior (solid) densities for the rate of glacial retreat at the *Zachariae Isstrøm* glacier.

Definition 3.1

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

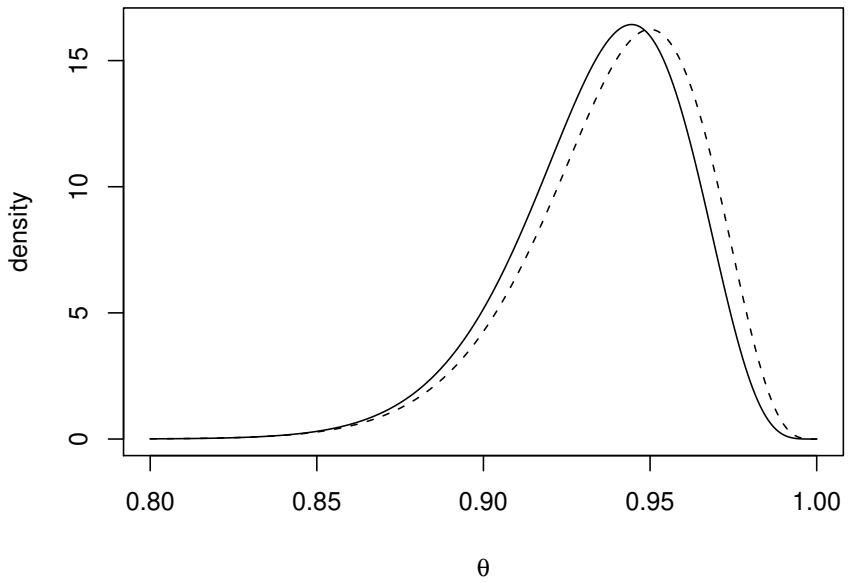

Figure 3.9: Prior (dashed) and posterior (solid) densities for θ

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 θ 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.

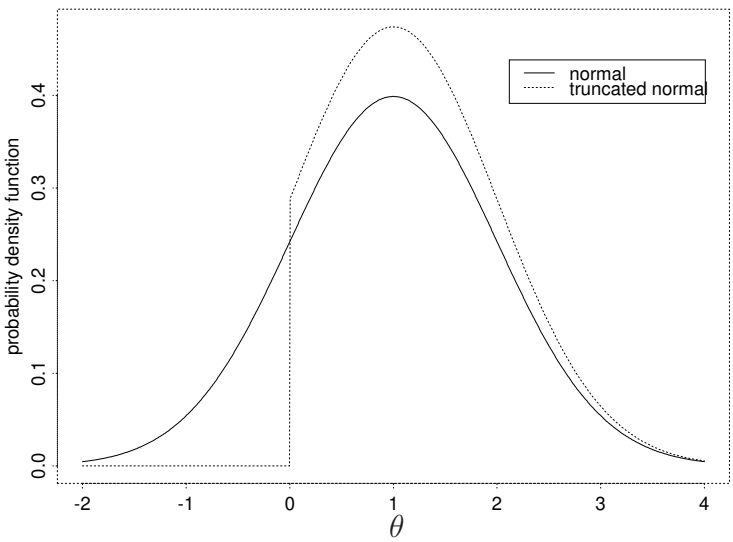

Figure 3.10: Plot of a normal and truncated normal distribution

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

Proof

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

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

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

Example 3.7

Consider again the case of $X_i|\theta \sim N(\theta, h^2)$, $i = 1, 2, \ldots, n$ (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 θ. Find the posterior distribution for θ.

...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.

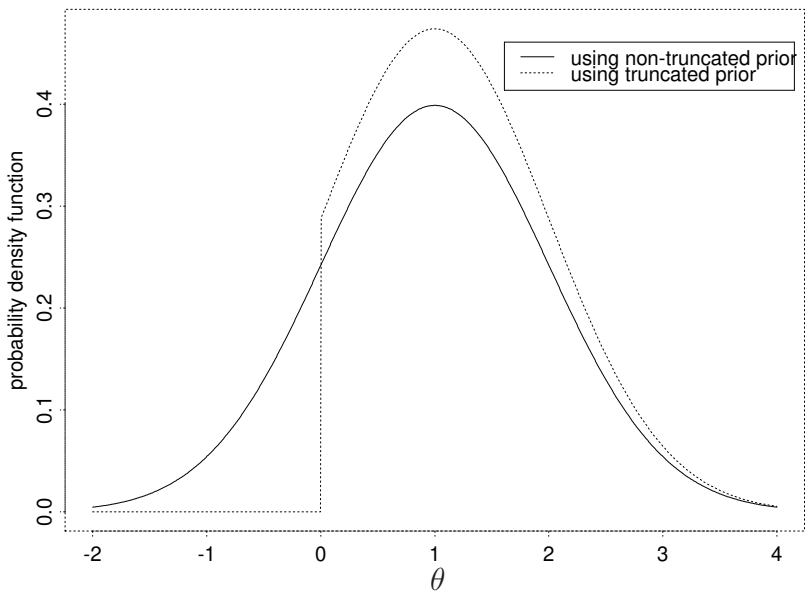

Figure 3.11: Plot of a posterior distribution determined using a non-truncated and a truncated prior distribution

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 θ. The prior probability that θ 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 θ 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 $\theta|\boldsymbol{x} \sim N(10, 0.25^2)$. 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.

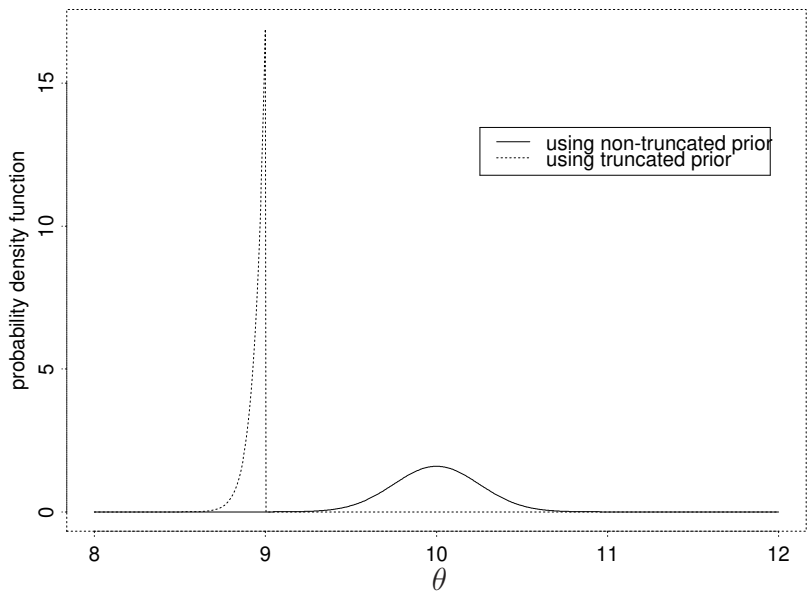

Figure 3.12: Plot of a posterior distribution determined using a non-truncated and a truncated prior distribution

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

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 $\theta =$ Pr(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 θ have come from observing the data.

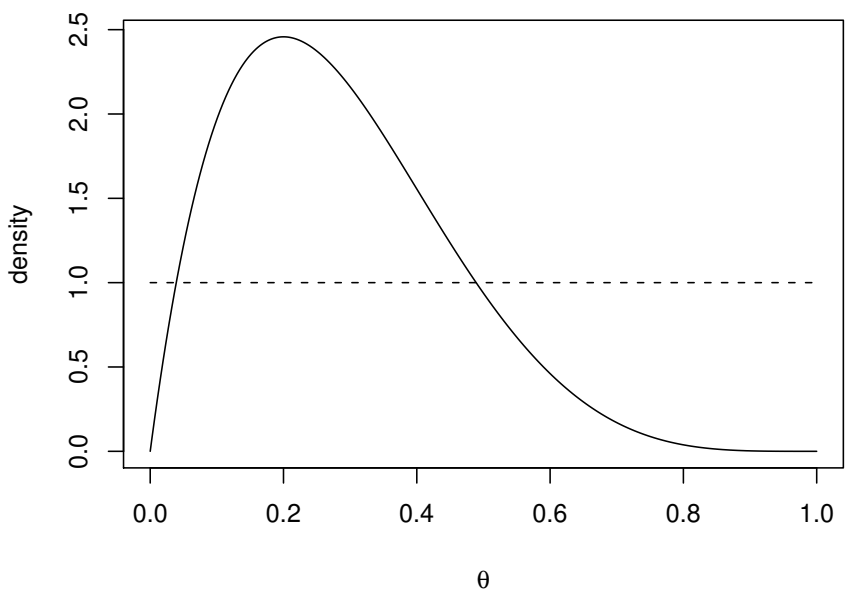

Figure 3.13: Prior (dashed) and posterior (solid) densities for θ

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 τ known). Determine the posterior distribution assuming a vague prior for μ.

✎ *...Solution to Example 3.9...*

Example 3.10

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

✎...*Solution to Example 3.10*...

Prior Ignorance

We could represent ignorance by the concept "all values of θ are equally likely". If θ were discrete with m possible values then we could assign each value the same probability $1/m$. However, if θ is continuous, we need some limiting argument (from the discrete case). Suppose that θ can take values between a and b, where $-\infty < a < b < \infty$. Letting all (permitted) values of θ be equally likely results in taking a uniform $U(a, b)$ distribution as our prior distribution for θ. 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

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

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 θ 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 ϕ: $\pi(\phi) = constant$. In turn, this induces a distribution on θ. 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

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

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

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

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 θ then we are also "ignorant" about any function of θ, 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 θ? If not, on which function of θ 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 θ. The above example shows that assigning a uniform/ignorance prior to θ means that we do not have a uniform/ignorance prior for e^{θ}.

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

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

He recommended that we represent prior ignorance by the prior distribution

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

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

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

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, $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (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

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

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 τ 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

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

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 μ and known variance σ^2), as the sample size $n \to \infty$. It is easy to show that $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$, and so

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

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 $n \to \infty$:

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

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

Theorem (Asymptotic posterior

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

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

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

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

Proof

Using Bayes Theorem, the posterior distribution for θ is

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

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

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

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

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

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

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

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

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

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

Hence

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

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

$$\sqrt{n}(\theta - \hat{\theta})|\boldsymbol{x} \xrightarrow{\mathcal{D}} N\left(0, [-\ell_n''(\hat{\theta})]^{-1}\right) \qquad \text{as } n \to \infty,$$

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

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

as required.

Comments

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

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

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 $\boldsymbol{X}|\theta$, that is, $I(\theta) = E_{\boldsymbol{X}|\theta}[J(\theta)]$.

3. This limiting result is similar to one for the maximum likelihood estimator in Frequentist Statistics:

$$\sqrt{I(\theta)}\ (\hat{\theta} - \theta)|\boldsymbol{x} \xrightarrow{\mathcal{D}} N(0,1). \qquad \text{as } n \to \infty, \tag{3.10}$$

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

Example 3.14

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

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

Determine the asymptotic posterior distribution for θ. Note that from Example 3.11 we have

$$\frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i^2,$$

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

...Solution to Example 3.14...

Example 3.15

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

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

✎ *...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 τ known). Determine the asymptotic posterior distribution for μ. Note that from Example 3.13 we have

$$\frac{\partial}{\partial \mu} \log f(\boldsymbol{x}|\mu) = n\tau(\bar{x} - \mu),$$
$$J(\mu) = -\frac{\partial^2}{\partial \mu^2} \log f(\boldsymbol{x}|\mu) = n\tau.$$

...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.

Chapter 4

Bayesian inference

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

4.1 Estimation

Point estimates

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

Interval estimates

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

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

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

A $100(1-\alpha)\%$ *highest density interval* (HDI) for θ is the region $C_\alpha = \{\theta : \ \pi(\theta|\boldsymbol{x}) \geq \gamma\}$ where γ is chosen so that $\Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$. This region is sometimes called a *most plausible Bayesian confidence interval*. If the posterior distribution has many modes then it is possible that the HDI will be the union of several disjoint regions; for example, the HDI could take the form $C_\alpha = (a,b) \cup (c,d) \cup (e,f)$, where $a < b < c < d < e < f$.

Interpretation of confidence intervals

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

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

Example 4.1

Suppose that the posterior distribution for θ is a $Beta(1, 24)$ distribution, with probability density function

$$\pi(\theta|\boldsymbol{x}) = 24\,(1-\theta)^{23}, \quad 0 < \theta < 1.$$

A plot of this distribution is given in Figure 4.1. Determine the $100(1-\alpha)\%$ HDI for θ.

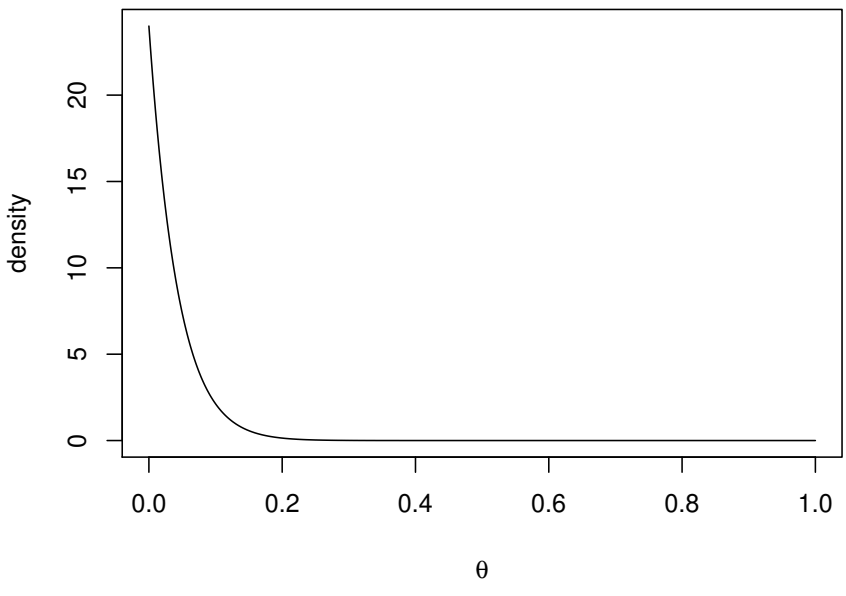

Figure 4.1: Plot of the $Beta(1, 24)$ posterior density function

✎ *...Solution to Example 4.1...*

Example 4.2

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

✎ *...Solution to Example 4.2...*

Example 4.3

Suppose that the posterior distribution for θ is a $Beta(2, 23)$ distribution, with probability density function

$$\pi(\theta|\boldsymbol{x}) = 552\,\theta(1-\theta)^{22}, \quad 0 < \theta < 1.$$

A plot of this distribution is given in Figure 4.2. Determine the $100(1-\alpha)\%$ HDI for θ.

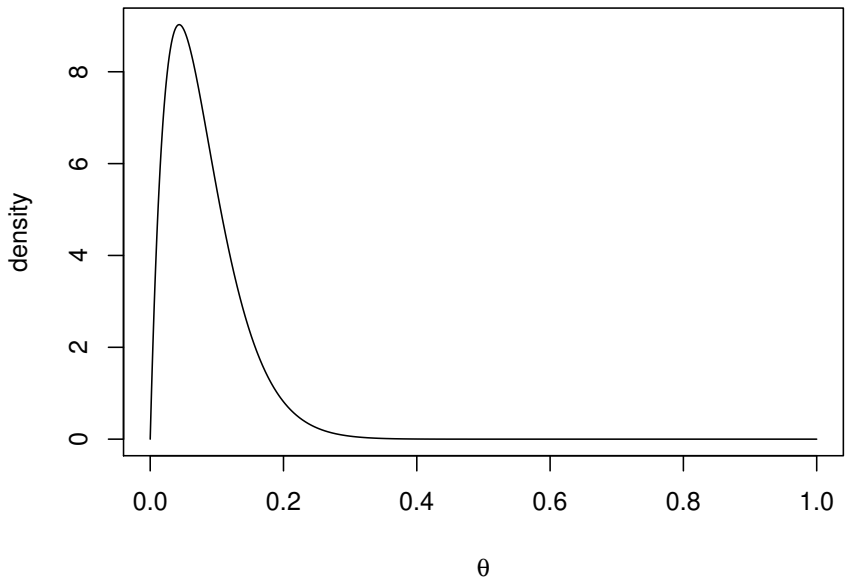

Figure 4.2: Plot of the $Beta(2, 23)$ posterior density function

...Solution to Example 4.3...

Computation of HDIs for unimodal distributions

Suppose that we require the HDI (a, b) for a unimodal distribution with density $f(\cdot)$ and distribution function $F(\cdot)$. We have seen that if one of the end-points is known (because of the shape of the distribution) or the distribution is symmetric then the solution is in terms of the distribution's percentage points. When this is not the case, the problem requires a numerical scheme to find a and b satisfying

$$F(b) - F(a) = 1 - \alpha \qquad \text{and} \qquad f(a) = f(b).$$

The solution can be found by noticing that it also minimizes the function

$$g(a, b) = \big\{F(b) - F(a) - (1 - \alpha)\big\}^2 + k\big\{f(b) - f(a)\big\}^2,$$

where $k > 0$ is a tuning parameter that tries to ensure that both terms are zeroed. Therefore, we can used the `R` optimizer function `optim` to determine a and b.

Example 4.3 (continued)

Suppose we need the 95% HDI for θ when $\theta|\boldsymbol{x} \sim Beta(2, 23)$. One slight complication with using the above method to determine the HDI (a, b) is that both a and b are restricted to the unit interval. However, the `R` function `optim` has options for dealing with such cases. It also needs initial guesses at the values of a and b. Here we base these on the values of the 95% equi-tailed Bayesian confidence interval. We also take $k = 0.0001$ to balance the conditions to be zeroed.

The `R` code to determine a and b is

```
g=function(x)
{
a=x[1]
b=x[2]
(pbeta(b,2,23)-pbeta(a,2,23)-0.95)^2+0.0001*(dbeta(b,2,23)-dbeta(a,2,23))^2
}

initiala=qbeta(0.025,2,23)
initialb=qbeta(0.975,2,23)
res=optim(c(initiala,initialb),g,method="L-BFGS-B",lower=0,upper=1)
a=res$par[1]
b=res$par[2]
```

and gives $a = 0.002211733$ and $b = 0.1840109$, with $F(b) - F(a) = 0.9500041$ and $f(b) - f(a) = -0.004484121$. Thus the 95% HDI is $(0.002211733, 0.1840109)$.

4.2 Prediction

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

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

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

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

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

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

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

Example 4.4

Suppose that X is the number of expensive goods sold in a shop over 24 days. If θ is the expected number of sales per day then it may be plausible that $X|\theta \sim Po(24\,\theta)$. Also, suppose our prior distribution for θ is as given in Table 4.1.

	θ	$\pi(\theta)$
"great"	1/2	0.2
"good"	1/4	0.5
"poor"	1/8	0.3

Table 4.1: Prior distribution for θ

Clearly, we believe that the most likely value of θ is 1/4, indicating that we would expect around 6 expensive goods to be sold in any 24 day period. Suppose now that we observe

that $x = 10$ expensive goods were sold in the last 24 days. This will impact our beliefs about θ. We can calculate the posterior distribution for θ as follows. The likelihood term is

$$\Pr(X = 10|\theta) = \frac{(24\theta)^{10}e^{-24\theta}}{10!} = \begin{cases} 0.1048 & \text{if } \theta = 1/2 \\ 0.0413 & \text{if } \theta = 1/4 \\ 0.0008 & \text{if } \theta = 1/8 \end{cases}$$

and so, using Bayes Theorem

$$\begin{aligned}\pi(\theta = 1/2|x = 10) &= \frac{\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2)}{[\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2) + \Pr(X = 10|\theta = 1/4)\pi(\theta = 1/4) \\ &\qquad + \Pr(X = 10|\theta = 1/8)\pi(\theta = 1/8)]} \\ &= \frac{0.1048 \times 0.2}{0.1048 \times 0.2 + 0.0413 \times 0.5 + 0.0008 \times 0.3} \\ &= 0.501\end{aligned}$$

$$\pi(\theta = 1/4|x = 10) = \cdots = 0.493$$
$$\pi(\theta = 1/8|x = 10) = \cdots = 0.006.$$

Thus, the posterior distribution for θ is as shown in Table 4.2, with most likely value of θ now being 1/2, and standard deviation $SD(\theta|x = 10) = 0.126$.

	θ	$\pi(\theta\|x = 10)$
"great"	1/2	0.501
"good"	1/4	0.493
"poor"	1/8	0.006

Table 4.2: Posterior distribution for θ

Suppose now we want to predict the number of sales Y in the next 24 days. If there have been no changes in the sales process (no special advertising campaigns etc) then we can take $Y|\theta \sim Po(24\,\theta)$. Determine the predictive probability function for Y.

✎ *...Solution to Example 4.4...*

✎...*Solution to Example 4.4 continued...*

This probability can be compared with a more naive predictive probability calculated assuming that $\theta = \hat{\theta}$, the likelihood mode. Here $\hat{\theta} = 1/2$ and so $Y|\theta = \hat{\theta} \sim Po(24\hat{\theta}) \equiv Po(12)$, whence

$$\Pr(Y = 10|\theta = \hat{\theta}) = \frac{12^{10}e^{-12}}{10!} = 0.1048.$$

In the same manner, we can calculate the entire predictive distribution and naive predictive distribution; see Table 4.3.

y	correct $f(y\|x=10)$	naive $f(y\|\theta=\hat{\theta})$
0	0.002	0.000
1	0.008	0.000
2	0.024	0.000
3	0.046	0.002
4	0.070	0.005
5	0.086	0.013
6	0.092	0.025
7	0.090	0.044
8	0.084	0.066
9	0.078	0.087
10	0.073	0.105
11	0.068	0.114
12	0.063	0.114
13	0.055	0.106
14	0.046	0.090
15	0.037	0.072
16	0.027	0.054
17	0.019	0.038
18	0.013	0.026
19	0.008	0.016
20	0.005	0.010
$\vdots$	$\vdots$	$\vdots$

Table 4.3: Predictive and "naive" predictive probability functions

Notice that the correct predictive probability distribution has more probability out in the tails of its distribution, that is, the probabilities of 0, 1, 2, ... are larger than their "naive" equivalents. This is a common occurrence. It is due to ignoring the uncertainty about the parameter estimate. Essentially, the naive predictive distribution is a predictive distribution which, instead of using the correct posterior distribution, uses the degenerate posterior distribution

$$\pi^*(\theta|x=10) = \begin{cases} 1 & \text{if } \theta = 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

a distribution with standard deviation $SD_{\pi^*}(\theta|x = 10) = 0$. The correct posterior standard deviation of θ is $SD_\pi(\theta|x = 10) = 0.126$. Therefore, the predictive distribution using the naive posterior π^* is, loosely speaking, too confident that it "knows" the value of θ and so produces a predictive distribution with too small a standard deviation:

$$SD(Y|x=10) = \begin{cases} 4.26 & \text{using the correct } \pi(\theta|x=10) \\ 3.46 & \text{using the naive } \pi^*(\theta|x=10). \end{cases}$$

These standard deviations can be calculated from Table 4.3.

We can also use the above table of predictive probabilities to determine a prediction interval for Y. Using the correct predictive distribution, and recalling the highest density feature of prediction intervals, we obtain $\Pr(2 \leq Y \leq 17|x = 10) = 0.959$. The corresponding naive calculation is $\Pr(6 \leq Y \leq 19|\theta = \hat{\theta}) = 0.958$. Hence the correct (approximate) 96% prediction interval for Y is $\{2, 3, \ldots, 17\}$. The naive version, and hence narrower interval, is $\{6, 7, \ldots, 19\}$.

Definition 4.1

The random variable Y follows a Beta-binomial $BetaBin(n, a, b)$ distribution (n positive integer, $a > 0$, $b > 0$) if it has probability function

$$f(y|n, a, b) = \binom{n}{y} \frac{B(y + a, b + n - y)}{B(a, b)}, \qquad y = 0, 1, \ldots, n,$$

where $B(a, b)$ is the beta function defined in (2.2). It can be shown that

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

Example 4.5

Suppose that X is the number of defective items in a sample of size 5. If the items are defective independently of one another and they each have the same probability θ of being defective then $X|\theta \sim Bin(5, \theta)$. Suppose we believe that defective items are quite unlikely and so take a $Beta(1, 19)$ prior distribution with mean and standard deviation

$$E(\theta) = 0.05 \qquad \text{and} \qquad SD(\theta) = 0.048.$$

Suppose we take a sample of size 5 and observe $x = 1$ defective item. In this case, the likelihood mode is $\hat{\theta} = 1/5 = 0.2$, higher than the prior mean. We have seen previously that, in such cases, the posterior distribution is a $Beta$ distribution whose first and second parameters are those of the prior distribution incremented by the number of success and the number of failures respectively. Thus, the posterior distribution is a $Beta(2, 23)$ distribution, with mean and standard deviation

$$E(\theta|x = 1) = 0.08 \qquad \text{and} \qquad SD(\theta|x = 1) = 0.053.$$

The posterior mean is larger than the prior mean and the standard deviation has also increased (slightly).

If we observe another sample of 5 items, what is the predictive probability distribution of the number found to be defective?

✎ *...Solution to Example 4.5...*

We can compare this predictive distribution with a naive predictive distribution based on an estimate of θ, for example, the likelihood mode or the posterior mode. Here we shall base our naive predictive distribution on the posterior mode $\hat{\theta} = 1/23$, that is, use the distribution $Y|\theta = \hat{\theta} \sim Bin(5, 1/23)$. Thus, the naive predictive probability function is, for $y = 0, 1, \ldots, 5$,

$$f(y|\theta = \hat{\theta}) = \binom{5}{y} \hat{\theta}^y (1 - \hat{\theta})^{5-y} = \binom{5}{y} \frac{22^{5-y}}{23^5}.$$

Numerical values for the predictive and naive predictive probability functions are given in Table 4.4. Again, the naive predictive distribution is a predictive distribution which,

y	correct $f(y\|x=1)$	naive $f(y\|\theta=\hat{\theta})$
0	0.680	0.801
1	0.252	0.182
2	0.058	0.017
3	0.009	0.001
4	0.001	0.000
5	0.000	0.000

Table 4.4: Predictive and naive predictive probability functions

instead of using the correct posterior distribution, uses a degenerate posterior distribution $\pi^*(\theta|x=1)$ which essentially allows only one value: $\Pr_{\pi^*}(\theta = 1/23|x=1) = 1$ and standard deviation $SD_{\pi^*}(\theta|x=1) = 0$. Note that the correct posterior standard deviation of θ is $SD_\pi(\theta|x=1) = 0.053$. Using a degenerate posterior distribution results in the naive predictive distribution having too small a standard deviation:

$$SD(Y|x=1) = \begin{cases} 0.652 & \text{using the correct } \pi(\theta|x=1) \\ 0.456 & \text{using the naive } \pi^*(\theta|x=1), \end{cases}$$

these values being calculated from $BetaBin(5,2,23)$ and binomial $Bin(5,1/23)$ distributions.

Using the numerical table of predictive probabilities, we can see that $\{0,1\}$ is a 93.2% prediction set/interval. This is to be contrasted with the more "optimistic" calculation using the naive predictive distribution which shows that $\{0,1\}$ is a 98.3% prediction set/interval.

Predictive distribution

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

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

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

Rearranging, we obtain

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

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

Reworking the previous example using this formula, we have

$$\theta \sim Beta(1, 19), \quad X|\theta \sim Bin(5, \theta), \quad Y|\theta \sim Bin(5, \theta)$$

from which we obtain

$$\theta|x = 1 \sim Beta(2, 23), \quad \theta|x = 1, y \sim Beta(y + 2, 28 - y).$$

Therefore, for $y = 0, 1, 2, \ldots, 5$

$$\begin{aligned} f(y|x = 1) &= \frac{f(y|\theta)\pi(\theta|x = 1)}{\pi(\theta|x = 1, y)} \\ &= \frac{\binom{5}{y}\theta^y(1-\theta)^{5-y} \times \dfrac{\theta(1-\theta)^{22}}{B(2,23)}}{\dfrac{\theta^{y+1}(1-\theta)^{27-y}}{B(y+2, 28-y)}} \\ &= \binom{5}{y}\frac{B(y+2, 28-y)}{B(2,23)}. \end{aligned}$$

Definition 4.2

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

$$f(y|a, b, c) = \frac{c^b y^{a-1}}{B(a, b)(y + c)^{a+b}} \qquad y > 0,$$

where $B(a, b)$ is the beta function defined in (2.2). It can be shown that

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

The distribution gets its name because $Y/(Y + c) \sim Beta(a, b)$. Also note that if $Y \sim InBe(a, b, 1)$ then $cY \sim InBe(a, b, c)$.

Example 4.6

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $Ga(k, \theta)$ distribution, where k is known, and our prior beliefs are described by a $Ga(g, h)$ distribution. The likelihood function is

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^k x_i^{k-1} e^{-\theta x_i}}{\Gamma(k)} \propto \theta^{nk} e^{-n\bar{x}\theta}.$$

Therefore, using Bayes Theorem, the posterior density is

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\
&\propto \theta^{g-1} e^{-h\theta} \times \theta^{nk} e^{-n\bar{x}\theta}, \quad \theta > 0 \\
&\propto \theta^{g+nk-1} e^{-(h+n\bar{x})\theta}, \quad \theta > 0.
\end{aligned}$$

Hence, the posterior distribution is a $Ga(G = g + nk, H = h + n\bar{x})$ distribution. Notice that this implies that the gamma distribution is the conjugate prior distribution for the model "random sample from a $Ga(k, \theta)$ distribution, with k known". Determine the predictive distribution for a future outcome Y.

...Solution to Example 4.6...

✎ *...Solution to Example 4.6 continued...*

Consider the case where the data follow an exponential distribution, that is, where $k = 1$. Determine the predictive density function and the $100(1-\alpha)\%$ prediction interval for Y.

✎ *...Solution to Example 4.6 continued...*

The End

Chapter 5

Exercises

Assignment questions

You will be told in class (and via email) which questions to answer for each of the assignments. Solutions to these questions should be placed in the homework submission letterbox in the foyer of the Maths & Stats General Office (Herschel Building 3rd floor) by no later than 4pm on the due date of the assignments.

Computing questions

This course will make some use of the statistical computing package R, and there will be occassional computing labs. You should bring this booklet to all lectures and computing labs. In the labs, you will be told which questions to work through. Some of these computing questions will be used for the assignments.

Non–assessed questions

We will work through most of the non–assessed questions in the problems classes; solutions to *all* questions will be made available through Blackboard.

Pencil-and-paper exercises

1. (a) Mark on a $0-1$ probability scale the approximate positions of the following events. Undertake a simple investigation if necessary. Explain the values you choose and state, in each case, whether you are using the frequency, classical or subjective interpretation of probability.

   (i) A student taking MAS2903 has a body mass index (BMI) of at least 22.5.

   (ii) Newcastle United will finish in a top 6 position in the Premier League this year.

   (iii) A randomly selected number between 10 and 99 (inclusive) is divisible by 7.

   (b) For each event in part (a), give a range which you believe contains the 'true' value with probability 0.95.

2. An insurance company classifies car drives as X, Y or Z. Experience indicates that the probability that a class X driver has at least one accident in any one year is 0.02, whilst the corresponding probabilities for class Y and class Z drivers are 0.04 and 0.10 respectively. They have also found that, of the drivers who apply to them for cover, 30% are class X, 60% are class Y and 10% are class Z.

   (i) Mr Smith is a new client and, within 12 months, he has an accident. What is the probability that he is a class Z risk?

   (ii) If Mrs Brown goes for n years without an accident, and we assume the number of accidents in different years to be independent, how large must n be before the company considers that she is more likely to belong to class X than to class Y?

3. Suppose that 1% of the population have a disease D. A diagnostic test S, designed to detect the disease, has the following accuracy:

$$\begin{aligned} \Pr(S \text{ positive}|D) &= 0.95 \qquad \text{and} \\ \Pr(S \text{ positive}|D^c) &= 0.05. \end{aligned}$$

   If 100 people were tested at random, how many would we expect to test positive and of those that do, about how many would we expect to have the disease?

4. On June 1st 2009, Air France flight 447 disappeared en-route to Paris Charles de Gaulle Airport about four hours after leaving Rio de Janeiro. Locating the aircraft's flight data recorder (the *black box*) was key to understanding the cause of one of the deadliest accidents in the history of Air France.

   After considering various oceanographic characteristics in the Atlantic at the time the flight disappeared, initial searches for the data recorder focussed on a large area to the north of the point at which it was believed the aircraft made impact with the ocean. In fact, it was believed that there was an 80% chance of finding the data recorder in this area, as opposed to the south of the impact zone. After almost two years of searching in the area mainly to the north of the point of impact, the elusive black box had still not been found.

   In April 2011 *Metron, Inc.* were hired to launch a Bayesian review of the search effort. The following probabilities of black box "identifiers" were estimated, based on the black box being in the areas defined as "North" and "South":

	Identifier			
	x_1	x_2	x_3	x_4
North	0.3	0.1	0.1	0.5
South	0.05	0.7	0.2	0.05

   These "identifiers" were:

   x_1 : "Fresh" debris found (since December 2010)
   x_2 : Signal detected from the black boxes
   x_3 : Both "fresh" debris found *and* signal from black boxes detected
   x_4 : Neither "fresh" debris found *nor* signals detected from the black boxes

   (a) Construct rules for differential diagnosis of the most plausible search area for the flight's data recorder, on the basis of the posterior assessments for each type of identifier.

   (b) According to your set of rules in part (a), what is the probability of 'misdiagnosis'?

   (c) In March 2011 a weak signal from the black box was detected. Quantify the shift in odds in favour of searching in the area south of the impact zone, having observed this identifier.

5. An experiment can result in four possible outcomes x_1, x_2, x_3 and x_4. The probability of each of these outcomes is affected by a parameter θ which can take one of three values θ_1, θ_2 and θ_3:

$p(x\|\theta)$	x_1	x_2	x_3	x_4
θ_1	0.1	0.2	0.3	0.4
θ_2	0.3	0.3	0.2	0.2
θ_3	0.5	0.1	0.1	0.3

The prior plausibilities attaching to θ_1, θ_2 and θ_3 are 0.1, 0.4 and 0.5. Find the posterior probability function for θ after observing x_1. Repeat the calculations for x_2, x_3 and x_4.

6. Extensive manuscripts of two scribes show that they differ on the relative frequencies with which they use two alternative vowel forms *e* and *oe*, as in the following table:

	e	*oe*
Scribe A	0.3	0.7
Scribe B	0.5	0.5

A new manuscript fragment, known to be by one of the scribes, uses the vowel form five times, three times as *e* and twice as *oe*. If an independent assessment of similar fragments favours Scribe A as author with probability 0.7, show that the evidence from the vowel form is sufficient to turn the odds in favour of Scribe B.

7. If $X_1, X_2, \ldots, X_n$ are independent random variables with probability function

$$f(x|\theta) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots,$$

show that $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

8. If $X_1, X_2, \ldots, X_n$ are independent $N(5, \sigma^2)$ random variables, show that

$$T = \sum_{i=1}^{n} (X_i - 5)^2$$

is sufficient for σ.

9. Hurricane–strength wind speeds (X miles per hour) for locations in the Gulf of Mexico are modelled using a distribution with probability density function

$$f(x|\sigma) = \frac{x}{\sigma^2} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}, \quad x > 0, \quad \sigma > 0.$$

During a hurricane, you collect wind speeds at n randomly chosen locations in the Gulf of Mexico, giving $x_1, x_2, \ldots, x_n$. Find the likelihood function for σ, and obtain a sufficient statistic for σ.

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

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

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

(a) obtain the posterior distribution, and compare this to the prior;

(b) find the posterior probability that $5.45\,g/cm^3 < \mu < 5.55\,g/cm^3$.

11. Suppose that a random sample $x_1, x_2, \ldots, x_n$ is obtained. Derive the posterior distributions for the following models:

(a) $f(x|\theta) = \theta^{x-1}(1-\theta)$, $x = 1, 2, \ldots$ with a $U(0,1)$ prior distribution for θ.

(b) $f(x|\theta) = \dfrac{e^{-\theta}\theta^x}{x!}$, $x = 0, 1, \ldots$ with a $Exp(1)$ prior distribution for θ.

(c) $f(x|\theta) = \dbinom{7}{x}\theta^x(1-\theta)^{7-x}$, $x = 0, 1, \ldots, 7$ with a $Beta(3,2)$ prior distribution for θ.

12. The time until a bicycle tyre gets punctured is exponentially distributed with mean $1/\theta$ weeks. From a cyclist, we are able to elicit the following prior for θ:

$$\pi(\theta) = \frac{100^6}{120}\theta^5 e^{-100\theta}. \tag{5.1}$$

(a) Identify fully the distribution in Equation (1).

(b) The cyclist records the times that two tyres last, and his readings (in weeks) are 33 and 41. Find how his beliefs about θ change in light of these data.

13. Consider an experiment with a possibly loaded six-sided die. Let $\theta = \Pr(\text{rolling a six})$. Before conducting the experiment, we believe that all values of θ are equally likely and so we use a $U(0,1)$ random variable as the prior distribution for θ. We then roll the die 10 times and observe 3 sixes.

(a) Obtain the likelihood function and use it to obtain the posterior distribution, $\pi(\theta|x = 3)$.

(b) What is $E[\theta|x = 3]$?

14. *[It might help if you have the handout for case study 1 handy when completing this question]*
Suppose the number of casualties Y_i at location i, due to road traffic accidents, is assumed Poisson with rate θ_i, i.e.

$$Y_i|\theta_i \sim Po(\theta_i).$$

Suppose further that we assume the following gamma prior for θ_i:

$$\theta_i \sim Ga\left(g, \frac{g}{\mu_i}\right).$$

(a) Show that the prior mean and standard deviation for θ_i are μ_i and $\mu_i/\sqrt{g}$ respectively.

(b) Suppose that we observe $Y_i = y_i$ casualties at site i during some observation period.

(i) Derive the posterior distribution $\pi(\theta_i|y_i)$, i.e. confirm the result given by Equation (3) in the case study.

(ii) Show that the posterior mean is given by the following Bayes linear rule:

$$E(\theta_i|y_i) = \alpha_i\mu_i + (1-\alpha_i)y_i,$$

where

$$\alpha_i = \frac{g}{g+\mu_i}.$$

15. Richard plays "flip to win" with a student. The game is very simple – Richard tosses a coin and if the outcome is a head, the student gets a fiver. If the coin lands tails, the student must buy Richard a pint. Five students play this game with Richard, and only one head is observed. Let X be the number of heads observed; thus, $X \sim bin(5, \theta)$.

(a) Find the likelihood function $f(x = 1|\theta)$.

(b) The students are suspicious of Richard's coin. In fact, Daniel, an insider, tells them that he believes the coin is biased and thus specifies a $Beta$(6,46) distribution as the prior for θ. Why might this be a sensible prior distribution in light of these suspicions?

(c) Obtain the posterior distribution for θ, $\pi(\theta|x = 1)$.

16. The number of defects in a 1200 foot roll of magnetic recording tape has a $Po(\theta)$ distribution. The prior distribution for θ is $Ga(3, 1)$. When 5 rolls of this tape are selected at random and inspected, the number of defects found on the rolls are 2, 2, 6, 0 and 3. Determine the posterior distribution of θ.

17. The data in the table below are measurements of the logarithm of *ornithine carbonyltransferase* (a liver enzyme) in patients suffering from acute viral hepatitis.

2.66	2.38	2.37	2.50	1.96	2.85	2.68	1.76	2.36
2.56	2.09	2.85	2.37	2.40	2.79	1.82	3.00	2.50
2.36	2.48	2.60	2.42	2.51	2.51	2.80	2.50	2.57

Assume that the enzyme measurement varies according to a $N(\mu, 0.27^2)$ distribution, and that an expert says his beliefs about the mean level μ are described by a $N(2.7, 0.3^2)$ distribution. After combining both sources of information, what is the probability that this mean level is less than 2.5? Sketch the prior and posterior distributions for μ on the same graph. Describe the main changes in beliefs after incorporating the data.

18. A random sample of size n is to be taken from a $N(\theta, 2^2)$ distribution. The prior for θ is $N(b, 1/d)$.

    (a) If $d = 1$, what is the smallest number of observations that must be included in the sample if the standard deviation of the posterior distribution of θ is to be reduced to 0.1?

    (b) If $n = 100$, show that no matter how large the value of d, the standard deviation of the posterior distribution is less than 1/5.

19. In viewing a section through the pancreas, doctors see what are called "islands". The numbers of islands seen in 900 patients are summarised below.

Number of islands	0	1	2	3	4	5	6	≥ 7
Frequency	327	340	160	53	16	3	1	0

Assuming that the numbers of islands are Poisson distributed with mean θ, obtain the posterior mean and standard deviation for θ from several gamma $Ga(g, h)$ prior distributions (with g and h in the range 1 to 50). Why is the posterior distribution quite insensitive to variations in the prior distribution?

20. Read through case study 2, and then answer the following questions.

    (a) Produce a list of benefits of working within the Bayesian framework over a standard frequentist analysis.

    (b) What difficulties arose as a result of working within the Bayesian framework?

    (c) How did we attempt to overcome these difficulties?

    (d) The resulting posterior distribution for θ was *improper*. What does this mean?

    (e) Name the simulation technique that can be used to obtain approximate draws from the posterior distribution when our prior is *non–conjugate*.

21. The *United States Geological Survey* (USGS) continuously tracks tropical depressions over the Atlantic Ocean – such systems can develop into hurricanes, and these storms can make landfall in eastern/southeastern states of the USA. You are a statistician working for the USGS, and in question 9 you have already used the *MATCH Uncertainty Elicitation Tool*, and R, to elicit a prior distribution for the rate of occurrence, θ, of three tropical depressions during the Atlantic hurricane season.

    Let X be the number of tropical depressions during the Atlantic hurricane season; we assume a Poisson distribution for X, i.e. $X \sim P(\theta)$. In 2010, 2011 and 2012, the USGS observe 15, 6 and 10 tropical depressions, respectively. Combine the prior opinion elicited from the meteorologist with the data to obtain the posterior distribution for θ, and write a short report for the meteorologist summarising how beliefs about the rate of occurrence of tropical depressions have changed in light of the data.

22. Suppose we have a random sample from a Poisson distribution, that is $X_i|\theta \sim P(\theta)$, $i = 1, \ldots, n$ (independent). Determine the posterior distribution assuming a vague prior for θ. What would be the resulting posterior if we had used a vague prior in question 21?

23. A "degassing burst" is a seismic event that can occur just before a major earthquake. Let θ be the chance that a major earthquake will occur after a degassing burst has taken place. From discussions with an seismologist, we are able to elicit the following:
$$\Pr(\theta < 0.6) = 10^{-3}.$$
    The seismologist also tells us that she thinks the most likely value for θ will be about 0.9. Use this information to justify using a $Beta(22.6, 3.4)$ prior for θ, clearly explaining your method. *[Hint: You may use R to help you.]*

24. Let θ be the chance that a major earthquake will occur after a "degassing burst" has taken place (see previous question). Six of the last ten degassing bursts at locations along the edge of the Eurasian plate were followed by major earthquakes. Update the beliefs of the seismologist from the previous question with this information to form the posterior for θ, and compare prior and posterior means.

25. The number of admissions, per day, to a busy Accident & Emergency (A&E) department is assumed *Poisson* with rate θ.

    This A&E department can cope with, at most, $\theta = 200$. The A&E consultant at this hospital tells us that he often sees between 80 and 100 patients admitted to A&E in any one day, although this can vary. Let $\Pr(80 < \theta < 100) = a$. We then elicit the following information from the consultant:

$$\begin{aligned}
\Pr(40 < \theta < 60) &\approx 0.5a; \\
\Pr(60 < \theta < 80) &\approx 0.8a \quad (= b); \\
\Pr(100 < \theta < 120) &\approx b; \\
\Pr(120 < \theta < 140) &\approx 0.2a \quad (= c); \\
\Pr(140 < \theta < 160) &\approx c; \qquad \text{and} \\
\Pr(160 < \theta < 180) &\approx 0.5c.
\end{aligned}$$

    In his time at this hospital, the consultant has never witnessed more than 180 casualties in any one day; only on very few occasions have there been less than 60 casualties.

    (a) Use the trial roulette method to obtain a suitable prior for θ, clearly demonstrating each step.

    (b) Is your prior in (a) *conjugate*? Explain.

    (c) Produce a feedback summary for the A&E consultant which gives:

    - the prior mode for θ;
    - the prior standard deviation for θ;
    - the 1% and 99% prior percentiles.

26. The number of casualties observed at a known traffic accident blackspot, in a single year, is assumed *Poisson* with rate λ. Using the bisection method, we are able to elicit the following quantities from a road safety expert:

$$Q_1(\lambda) = 2, \qquad Q_2(\lambda) = 4 \qquad \text{and} \qquad Q_3(\lambda) = 8,$$

    where Q_1, Q_2 and Q_3 are the first, second and third quartiles (respectively). The expert also tells us that the mean casualty rate at this site is extremely unlikely to exceed 20 per year.

    (a) Use *MATCH* to elicit an appropriate prior distribution for θ, demonstrating this clearly.

    (b) Given that we observed 6 casualties at this location in 2011, obtain the posterior distribution for λ.

27. The proportion θ of defective items in a large shipment is unknown. However, experience suggests that a $Beta(2, 200)$ prior is appropriate.

    (a) Suppose that 100 items are selected at random from the shipment and that three are found to be defective. What is the posterior distribution of θ?

    (b) Suppose that another statistician, having observed the three defectives, said that his posterior distribution for θ was a beta distribution with mean $4/102$ and variance 0.0003658. What prior distribution had that statistician used?

28. Suppose that the random variable Y follows a $Ga(a, b)$ distribution. Determine the probability density function for the random variable $X = 1/\sqrt{Y}$.

29. Using a random sample from a $Bin(k, \theta)$ (with k known), determine the posterior distribution for θ assuming

    (i) vague prior knowledge;

    (ii) the Jeffreys prior distribution;

    (iii) a very large sample.

30. Using a random sample from a $Ga(k, \theta)$ distribution (with k known), determine the posterior distribution for θ assuming

    (i) vague prior knowledge;

    (ii) the Jeffreys prior distribution;

    (iii) a very large sample.

31. Use techniques from MAS2901 to obtain 95% frequentist confidence intervals for θ_{C} and θ_{I} in question 9.

32. Compare the frequentist and Bayesian confidence intervals for θ_{C} and θ_{I} as obtained in questions 31 and 44 (respectively). Comment on any numerical differences, as well as differences in interpretation.

33. Sketch a probability density function for which the highest density interval (HDI) takes the form

    (i) (a, b)

    (ii) $(a, b) \cup (c, d)$

    (iii) $(a, b) \cup (c, d) \cup (e, f)$.

    where $a < b < c < d < e < f$.

34. The weights of items from a certain production processes are independently and identically distributed, each with a $N(\theta, 4)$ distribution. The production manager believes that θ varies from batch to batch according to a $N(110, 0.4)$ distribution. A sample of 5 items is randomly selected from a batch, giving data: 108, 109, 107.4, 109.6, 112. Determine the posterior distribution for θ.

    (i) Determine the predictive distribution for the weight of one further item from the batch.

    (ii) Calculate the (predictive) probability that the weight of this item exceeds 110.

    (iii) Determine the predictive distribution for the sample mean of the weights of m further items from the batch. What happens as $m \to \infty$?

    Hint: if the posterior distribution is $\theta|\boldsymbol{x} \sim N(B, 1/D)$ then the predictive distribution for a future item y, where $Y|\theta \sim N(\theta, 1/k)$, is

    $$f(y|\boldsymbol{x}) = \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)} = \left\{\frac{kD}{2\pi(D+k)}\right\}^{1/2} \exp\left\{-\frac{kD}{2(D+k)}(y-B)^2\right\},$$

    that is, $Y|\boldsymbol{x} \sim N\left(B, \dfrac{D+k}{kD}\right)$.

35. The distribution of flaws along the length of an artificial fibre follows a Poisson process, and the number of flaws in a length ℓ is $Po(\ell\theta)$. Very little is known about θ. The number of flaws in five fibres of lengths 10, 15, 25, 30 and 40 metres were found to be 3, 2, 7, 6, 10 respectively. We are interested in finding the predictive distribution for the number of flaws in another piece of length 60 metres.

    Consider first the general case.

    (i) Suppose the number of flaws are independent with $X_i \sim Po(\ell_i\theta)$. Determine the posterior distribution for θ.

    (ii) Determine the predictive distribution for the number of flaws Y in another piece of length ℓ.

    (iii) Use the information in the data to determine the number of flaws in another piece of length 60 metres.

    Hint for (ii): what is the probability function for $T = W + r$ when $W \sim NegBin(r, p)$?

Computing exercises

Open R by clicking on `Start` → `All Programs` → `Statistical Software` → `RStudio`. You will need to access data for some of these questions from the course website, so have this open:

`http://www.mas.ncl.ac.uk/~nlf8/teaching/mas2903/`

Some questions will also make use of the *MATCH Uncertainty Elicitation Tool*, available at:

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

36. Recall Example 2.1 in the lecture notes. We have an experiment with a possibly biased coin, and $\theta = \Pr(\text{Head})$. The number of heads observed, X, from 5 flips of the coin, is therefore a binomial random variable, i.e.

$$X|\theta \sim Bin(5, \theta).$$

Assuming that $\theta \sim U(0,1)$, we found that $\theta|x = 1 \sim Be(2,5)$. Let's now re–produce the plots in Figure 2.3, comparing prior with posterior.

In R, type the following:

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

You can now look at `x` by typing

```
> x
  0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13 0.14
  0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
  0.30 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44
  0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
  0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74
  0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89
  0.90 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```

Thus, `x` is just a sequence of values between 0 and 1, in steps of 0.01. This covers the range of possible for θ.

Now, for the $U(0,1)$ prior for θ, we apply the R function `dunif` to `x`:

```
> prior=dunif(x,0,1)
```

This evaluates the $U(0,1)$ pdf at each of our `x` values.

For the $Be(2,5)$ posterior for θ, we apply the R function `dbeta` to `x`:

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

This evaluates the $Be(2,5)$ pdf at each of our `x` values.

We will now plot the prior and posterior in R on the same graph, as in Figure 2.3 of the lecture notes. In R, type

```
> plot(x,prior,type='l',xlab='theta',main='Example 2.3',lty=2)
```

This should plot your $U(0,1)$ prior for θ. The arguments `xlab=` and `main=` specify labels for your x–axis and the plot as a whole, respectively. Using `type='l'` uses a line to join the points in the plot, and setting `lty=2` uses a dashed line effect.

Now to superimpose the posterior. Type

```
> lines(x,posterior)
```

Notice that when you do this, your beta posterior is superimposed (the `lines` command superimposes, instead of making a brand new plot); however, the range on the y–axis is not large enough to see this properly. We can fix this by using the command `ylim=c(lower,upper)` to specify a more suitable range. For example, inserting `ylim=c(0,max(posterior))` along with `type='l'`, `xlab='theta'` etc. within the plotting command `plot(...)` starts the plot again, but now the y–axis will go from 0 to the largest value in our posterior distribution. Now

```
> lines(x,posterior)
```

should produce the desired effect. Check this against Figure 2.3 in the lecture notes.

37. Recall Example 2.3 in the lecture notes. Max, the video game pirate, is interested in the proportion θ of customers who will buy *Call of Duty* from him. X is the number of people who say they *would* buy this game; if he asks 5 people, then we have

$$X|\theta \sim Bin(5,\theta).$$

He assumes a beta prior for θ: $\theta \sim Be(2.5, 12)$, which results in $\theta|x = 4 \sim Be(6.5, 13)$. Following carefully what you did in question 36, produce a plot in R which compares the prior and posterior distributions for θ. Compare your plot to that in Figure 2.7 of the lecture notes.

38. Recall the earthquakes example in the lecture notes: Example 2.4. Here, we assume the times between earthquakes X are exponentially distributed, i.e.

$$X|\theta \sim Exp(\theta).$$

An expert tells us that $\theta \sim Ga(10, 4000)$. We then combine our observations with the expert's prior beliefs to obtain $\theta|\boldsymbol{x} \sim Ga(30, 13633)$. Recall that our strong expert opinion suggest that only very small values of θ are likely (as earthquakes are relatively rare); thus, set up your x values in R by typing

```
> x=seq(0,0.01,0.001)
```

Of course, the gamma distribution is defined over the *entire* positive real line, but we only look at very small values of θ here. Now use the R command `dgamma` in exactly the same way you used `dunif` and `dbeta` in questions 36 and 37 to produce a plot comparing your prior and posterior distributions for the rate of earthquakes θ. You might have to change the range of the y–axis in your plot to accommodate the posterior. Compare you plot to that in Figure 2.9 of the lecture notes.

39. In R, type

```
> par(mfrow=c(2,2))
```

You are required to produce four plots in this question; the above command splits the plotting window into four components (2 by 2), one for each of the plots you will produce. You should include this single panel showing all four plots in your solutions to this question. Using the same approach as in question 36,

(a) produce a plot of the $Be(10, 37)$ density for values on the x–axis between 0 and 1;

(b) produce a plot of the $Be(20, 77)$ density for values on the x–axis between 0 and 1;

(c) produce a plot of the $Ga(15, 0.625)$ density for values on the x–axis between 0 and 100;

(d) produce a plot of the $Ga(9, 0.36)$ density for values on the x–axis between 0 and 100.

You should now have a single panel of four plots, one each for parts (a), (b), (c) and (d). Now read the following scenarios.

> Scenario 1
>
> A football fan tells us it is quite unlikely that her team will get relegated this year. Let $\theta = \Pr(\text{her team will be relegated})$. We are able to elicit from her that there is less than a 1 in 100,000 chance that $\theta > 0.5$.

> Scenario 2
>
> We are interested in the rate of retreat θ (feet per year) of the *Zachriae Isstrøm* glacier in Greenland, as this could be an indicator of climate change (see the recent BBC *Frozen Planet* series). An expert glaciologist tells us that this year, he expects this glacier to retreat by about 24 feet, although there is some uncertainty here. In fact, he thinks that there is 1/100 chance that the glacier might retreat by less than 10 feet.

(e) Match θ in each of the scenarios above with the most suitable distribution from parts (a)—(d). You can use the R commands `pbeta(q,a,b)` and `pgamma(q,a,b)` to find $\Pr(\theta < q)$, for $\theta \sim Be(a, b)$ and $\theta \sim Ga(a, b)$, respectively.

Questions 40 and 41 relate to **case study 1: speed cameras and regression to the mean**. Recall that there were 17 road traffic accident blackspot sites identified in South Tyneside (see Section 5 of the case study). The file `blackspots.txt` on the course website has data relating to the average observed speed of vehicles, and the number of casualties in the periods *before* and *after* the introduction of speed cameras, at each of these sites.

Recall also that Northumbria Police collected information on the average observed vehicle speed and the observed number of casualties (throughout the same time as the *before* period for the blackspot sites) from a further 67 "reference" sites; these data can be found on the course website in the file `reference.txt`.

40. This question relates to data collected at the **reference sites**.

    (a) Download the reference data from the MAS2903 course webpage by executing the following code in R:

    ```
    > ref = read.table("http://www.mas.ncl.ac.uk/~nlf8/teaching/
      + mas2903/case/reference.txt")
    ```

    Do not enter the + sign. I have done this because the code would not fit on a single line in this handout. You should enter the code in a single line!

    (b) Look at the data by typing

    ```
    > ref
    ```

    The first column contains the average observed speed of vehicles at each of the 67 reference sites, and the second column contains the number of casualties at each of these sites, during an observation period. Typing

    ```
    > AvSp = ref[,1]
    > Cas = ref[,2]
    ```

    will store the data in column 1 of `ref` in the object `AvSp`, and the data in column 2 of `ref` in the object `Cas`.

    (c) Use the `plot` command in R to produce a scatterplot of the number of casualties (Y) against the average observed speed (X). You should make sure you have appropriately labelled axes and a title on your plot; if in doubt, typing

    ```
    > ?plot
    ```

    will bring up a help screen for the command `plot`. Comment on the relationship you see, but at this stage *do not include this plot in your solutions.*

    (d) Perform a simple linear regression of casualties on the average observed speed by typing

    ```
    > lm(Cas~AvSp)
    ```

    In your solutions, write down the resulting regression equation, and make sure this confirms what you see in Equation (2) of the case study.

(e) You should now superimpose the regression line from part (d) onto your plot in part (c) and include this updated plot in your solutions. This can be done by typing:

```
> lines(AvSp,fitted(lm(Cas~AvSp)))
```

41. This question will use your solutions to question 40 and the equations given in the case study handout, as well as data collected at the **blackspot sites**.

(a) Some results based on analyses of blackspot sites 1 and 10 were given in the case study (see Section 5.5). Now, each of you have been randomly allocated to one of the other 15 blackspot sites; go to the "Case Studies" section of the course website to find out which site you have been allocated to, and make a note of this.

(b) Download the blackspot dataset in exactly the same way you downloaded the reference dataset in question 40(a), and store this in the object `blackspot` in R (as you stored the reference data in `ref`).

(c) Look at the data. As with the reference data, the first column relates to the average observed vehicle speed at each site. The second column relates to the number of casualties at each blackspot site in the period *before* the installation of speed cameras, and the third column gives the corresponding number in the period *after* the cameras had been installed. As before, each row is for one of the sites. Look at the row for the blackspot site you have been allocated to; in your solutions, write down your blackspot site number ($i =$), the average observed vehicle speed for your site ($x_i =$), the number of observed casualties at this site in the *before* period ($y_i =$) and the number of casualties in the *after* period ($y_{i,\text{after}} =$).

(d) Estimate the mean casualty frequency μ_i at your blackspot site by substituting your average observed vehicle speed x_i into the regression equation obtained in question 40(d) from the reference data. Make sure you write this in your solutions.

(e) In the case study, you are told that a road safety expert has suggested that $g = 2.5$ in the following distribution for the casualty rate θ_i at blackspot site i:

$$\theta_i \sim Ga\left(g, \frac{g}{\mu_i}\right)$$

(see question 14 and Equation (1) in the case study). Use your answer to part (d) to write down the distribution for θ_i for *your* blackspot site.

(f) Assuming that Y_i is the number of casualties at blackspot i, and that

$$Y_i|\theta_i \sim Po(\theta_i),$$

use your answer to question 14(b)(i), or Equation (3) in the case study, to write down the posterior distribution for $\theta_i|y_i$ for *your* blackspot site.

(g) Using similar commands in R to those used in questions 36–38, produce a plot of the prior and posterior distribution of θ_i for *your* blackspot site, similar to that shown for site 1 in Figure 4 of the case study. Make sure these two plots are given on the same graph (use the `lines` command in R for the second graph). Also, show the position of your observed casualty frequency by using the command

```
> abline(v=*)
```

where you would replace $*$ with *your* observed casualty frequency y_i.

(h) With reference to Equation (4) the case study, or question 14(b)(ii), calculate the posterior mean $E(\theta_i|y_i)$ for *your* blackspot site. Does this give more weight to the prior mean μ_i or the observed casualty frequency y_i?

(i) After accounting for the phenomenon of *Regression To the Mean* (RTM), compute the percentage reduction in casualty frequency attributed to the speed camera at *your* blackspot site.

42. The *United States Geological Survey* (USGS) continuously tracks tropical depressions over the Atlantic Ocean – such systems can develop into storm systems or hurricanes, and these storms can make landfall in eastern/southeastern states of the USA. You are a statistician working for the USGS, and you are interested in the rate of occurrence, θ, of these tropical depressions during the Atlantic hurricane season (June–November).

After a consultation with a meteorologist on the team, you find out that, in any year, she would probably expect to see between 16 and 20 tropical depressions over the Atlantic ocean. Let $\Pr(16 < \theta < 20) = a$.

You then ask her to specify how likely she thinks some other ranges for θ are, relative to the range she specified as most likely (given above). The results of your consultation are shown below.

$$\begin{aligned}
\Pr(12 < \theta < 16) &= 0.8a \\
\Pr(20 < \theta < 24) &= 0.8a \\
\Pr(8 < \theta < 12) &= 0.5a \\
\Pr(24 < \theta < 28) &= 0.2a \\
\Pr(28 < \theta < 32) &= 0.2a \\
\Pr(32 < \theta < 36) &= 0.1a
\end{aligned}$$

The meteorologist tells you that she has never seen more than 40 tropical depressions in any one year; she also tells you that during a "quiet" hurricane season there might no tropical depressions observed at all. However, she assigns values of θ outside the range specified above negligible probabilities.

(a) In *MATCH*, enter the `Lower Limit` and `Upper Limit` as specified by the meteorologist. Then select the `Roulette Input Mode` to perform a trial roulette

elicitation. A grid should appear; the x–axis should be split into equal–width *bins* in which you can place your *chips*.

(b) Place 10 chips in the 16→20 bin; this will represent the meteorologist's probability a. Now use the information you elicited from the meteorologist to place the remaining chips.

(c) Now click `Fitting & Feedback` on the toolbar at the top of the screen. In your solutions, copy and complete the following, specifying the parameters of your fitted distribution to two decimal places.

Feedback: first fit

$$\theta \sim$$

$$\pi(\theta) = \qquad\qquad\qquad\qquad\qquad\qquad , \qquad \theta > 0$$

$$E(\theta) =$$

$$SD(\theta) =$$

(d) Use the `Fitting & Feedback` box in *MATCH* to find the 1st and 99th percentiles of your prior distribution, and write these values down as probability statements in your solutions.

(e) Phrase a question that you could ask the meteorologist using your answers to part(d), as part of the feedback stage necessary in any elicitation process.

(f) In response to your question in part (e), the meteorologist would like you to refine the fitted prior for θ; in particular, she believes the 99th percentile is too high, and that

$$\Pr(\theta > 32) \leq 0.01.$$

In response to the meteorologist, remove a single chip from bins $24 \to 28$ and $28 \to 32$. Also remove two chips from bin $20 \to 24$. You should see that the meteorologist's request has now been met. Now provide her with a new summary sheet, by copying and completing the following in your solutions (see overleaf):

Feedback after refinement: 2nd fit

$$\theta \sim$$

$$\pi(\theta) = \qquad\qquad\qquad\qquad\qquad\qquad , \qquad \theta > 0$$

$$E(\theta) =$$

$$SD(\theta) =$$

$$\Pr(\theta < _____) = 0.01$$
$$\Pr(\theta > _____) = 0.01.$$

(g) Now use R:

(i) To produce a plot of $\pi(\theta)$ obtained from your second fit in *MATCH*, and include this in your solutions. *[Hint: you may find it useful to refer to questions 36–38]*

(ii) To confirm the feedback percentiles you obtained from your second fit in *MATCH*, and include any relevant R code in your solutions.

43. In this question we return to the scenario in question 21. Suppose the USGS are also interested in the probability, λ, that a tropical depression over the Atlantic develops into a major hurricane. After another consultation with the meteorologist, we are able to elicit that the mode of any distribution for λ should be around 0.7 and that $\Pr(\lambda < 0.3) = 0.001$.

(a) From the following list, choose the most appropriate distribution for λ: $U(a, b)$, $N(a, b)$, $Be(a, b)$, $Ga(a, b)$.

(b) For the model you selected in (a), write down an expression for the mode, and use the information elicited from the meteorologist to find an expression for a in terms of b.

(c) Now write a function in R like that at the top of page 54 in the lecture notes, which evaluates the cumulative distribution function for your chosen distribution, minus 0.001, at b.

(d) Now use the R command `uniroot`, as we do on page 54 of the lecture notes, to find b.

(e) Now find a, and write down your elicited prior for λ.

44. Gravestones provide a useful means of observing and measuring the degradation of different types of rock over time. One method of measuring the extent of degradation is *Lettering Alteration*, as outlined in Rahn (1971) and Meierdin (1981).

    In a recent study, researchers were interested in the degradation of granite. To investigate, a random sample of granite gravestones was taken from a cemetery in Savannah, a city along the southeast coast of the USA; the method of lettering alteration was used to assess the degradation of these gravestones. These data can be found in the file `granite-coastal` on the course website.

    A random sample of granite gravestones was also taken from a cemetery in Macon, Georgia, a town located 170 miles inland from Savannah. These data can be found in the file `granite-inland` on the course website.

(a) Download the `granite-coastal.txt` dataset into R by typing:

```
>coastal=read.table('http://www.mas.ncl.ac.uk/~nlf8/teaching/
+  mas2903/problems/granite-coastal.txt')
```

Make sure you do something similar for the `granite-inland.txt` dataset. You can the look at the data by typing `coastal` and `inland` at the R command prompt. Notice that each dataset contains two columns: for each gravestone, we have the age of the gravestone (the difference between the current year and the year the person died), and the depth of degradation as determined by the lettering alteration method. For each dataset, you should obtain the rate of degradation per year, by dividing the depth of degradation by the age of the gravestone. For example, for the coastal site:

```
> rate.coastal=coastal[,2]/coastal[,1]
```

Use R to find the mean and standard deviation annual rate of granite degradation for each dataset, and show these values in your solutions.

(b) Researchers are interested in rock which degrades at a rate of more than 0.12mm per year. From your data in R, estimate θ_C and θ_I, the proportion of gravestones with an annual degradation rate of at least 0.12mm at the coastal and inland locations, respectively. Also find the likelihood functions for θ_C and θ_I.

(c) A geologist is asked about the proportion of granite gravestones she would expect to see with a rate of degradation of at least 0.12mm per year. The geologist explains that gravestones in coastal locations are prone to more severe

degradation due to the salt content in the air; the table below summarises the expert's beliefs about θ_C and θ_I. An entry of ε indicates that the geologist thinks this range for θ is extremely unlikely.

θ	Coastal	Inland
$0 \to 0.1$	ε	a
$0.1 \to 0.2$	ε	$b = 0.4a$
$0.2 \to 0.3$	ε	b
$0.3 \to 0.4$	ε	ε
$0.4 \to 0.5$	ε	ε
$0.5 \to 0.6$	$d = 0.1c$	ε
$0.6 \to 0.7$	$2d$	ε
$0.7 \to 0.8$	$4d$	ε
$0.8 \to 0.9$	c	ε
$0.9 \to 1$	$8d$	ε

(i) Given this information from the expert, which method of prior elicitation might be suitable to obtain prior distributions for θ_C and θ_I?

(ii) Use this method to obtain $\pi(\theta_C)$ and $\pi(\theta_I)$ (rounding the parameters of these priors to the nearest integer).

(iii) Briefly explain how these prior distributions reflect the expert's beliefs about the differences between θ_C and θ_I.

(d) Now combine the geologist's beliefs with the data to obtain the posterior distributions for θ_C and θ_I.

(e) Use R to obtain plots of the prior and posterior for θ_C and θ_I, superimposing the prior and posterior on the same graphs in each case. Compare and contrast:

(i) Prior beliefs with posterior beliefs, for both θ_C and θ_I;

(ii) Posterior beliefs about the rate of granite degradation at the coast and inland.

(f) Obtain the 95% highest density interval for θ_I. *[Hint: You may use the results from Example 4.1 in the lecture notes to help you here.]*

(g) Why is the 95% highest density interval for θ_C more awkward to find? Use the R function `optim` to obtain this interval, and include your R code in your solutions.

(h) In light of the data, would you say there is a significant difference in the rate of granite degradation between coastal and inland locations? Explain your answer.

References

Meierdin, T.C. (1981). Marble weathering rates: a transect of the United States. *Physical Geography*, **2**, pp. 1—18.

Rahn, T. (1971). The weathering of tombstones and its relation to the topography of New England. *Journnal of Geological Education*, **19**, pp, 112—118.