CHAPTER 1

Bayesian Basics Part 1

1.1 Introduction

Bayesian methods is a term which may be used to refer to any mathematical tools that are useful and relevant in some way to *Bayesian inference*, an approach to statistics based on the work of Thomas Bayes (1701–1761). Bayes was an English mathematician and Presbyterian minister who is best known for having formulated a basic version of the well-known *Bayes' Theorem*.

Figure 1.1 (page 3) shows part of the Wikipedia article for Thomas Bayes. Bayes' ideas were later developed and generalised by many others, most notably the French mathematician Pierre-Simon Laplace (1749–1827) and the British astronomer Harold Jeffreys (1891–1989).

Bayesian inference is different to *classical inference* (or *frequentist inference*) mainly in that it treats model parameters as *random variables* rather than as *constants*. The Bayesian framework (or paradigm) allows for prior information to be formally taken into account. It can also be useful for formulating a complicated statistical model that presents a challenge to classical methods.

One drawback of Bayesian inference is that it invariably requires a prior distribution to be specified, even in the absence of any prior information. However, suitable *uninformative* prior distributions (also known as *noninformative*, *objective* or *reference* priors) have been developed which address this issue, and in many cases a nice feature of Bayesian inference is that these priors lead to exactly the same point and interval estimates as does classical inference. The issue becomes even less important when there is at least a moderate amount of data available. As sample size increases, the Bayesian approach typically converges to the same inferential results, irrespective of the specified prior distribution.

Another issue with Bayesian inference is that, although it may easily lead to suitable formulations of a challenging statistical problem, the types of calculation needed for inference can themselves be very complicated. Often, these calculations take on the form of multiple integrals (or summations) which are intractable and difficult (or impossible) to solve, even with the aid of advanced numerical techniques.

In such situations, the desired solutions can typically be approximated to any degree of precision using *Monte Carlo* (MC) methods. The idea is to make clever use of a large sample of values generated from a suitable probability distribution.

How to generate this sample presents another problem, but one which can typically be solved easily via *Markov chain Monte Carlo* (MCMC) *methods*. Both MC and MCMC methods will feature in later chapters of the course.

1.2 Bayes' rule

The starting point for Bayesian inference is *Bayes' rule*. The simplest form of this is

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\bar{A})P(B \mid \bar{A})},$$

where A and B are events such that $P(B) > 0$. This is easily proven by considering that:

$$P(A \mid B) = \frac{P(AB)}{P(B)} \quad \text{by the definition of conditional probability}$$

$$P(AB) = P(A)P(B \mid A) \quad \text{by the multiplicative law of probability}$$

$$P(B) = P(AB) + P(\bar{A}B) = P(A)P(B \mid A) + P(\bar{A})P(B \mid \bar{A})$$

by the law of total probability.

We see that the posterior probability $P(A \mid B)$ is equal to the prior probability $P(A)$ multiplied by a factor, where this factor is given by $P(B \mid A) / P(B)$.

As regards terminology, we call $P(A)$ the *prior* probability of A (meaning the probability of A *before* B is known to have occurred), and we call $P(A \mid B)$ the *posterior* probability of A *given* B (meaning the probability of A *after* B is known to have occurred). We may also say that $P(A)$ represents our *a priori* beliefs regarding A, and $P(A \mid B)$ represents our *a posteriori* beliefs regarding A.

Figure 1.1 Beginning of the Wikipedia article on Thomas Bayes

Source: en.wikipedia.org/wiki/Thomas_Bayes, 29/10/2014

Thomas Bayes

From Wikipedia, the free encyclopedia

Thomas Bayes (/ˈbeɪz/; c. 1701 – 7 April 1761)[1][2][note a] was an English statistician, philosopher and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name: Bayes' theorem. Bayes never published what would eventually become his most famous accomplishment; his notes were edited and published after his death by Richard Price.[3]

Thomas Bayes

Portrait used of Bayes in the 1936 book *History of Life Insurance*; it is dubious whether it actually depicts Bayes.[1] No earlier portrait or claimed portrait survived.

Born	c. 1701 London, England
Died	7 April 1761 (aged 59) Tunbridge Wells, Kent, England
Residence	Tunbridge Wells, Kent, England
Nationality	English

Signature

T. Bayes.

Contents [hide]

Biography [edit]

Thomas Bayes was the son of London Presbyterian minister Joshua Bayes,[4] and was possibly born in Hertfordshire.[5] He came from a prominent nonconformist family from Sheffield. In 1719, he enrolled at the University of Edinburgh to study logic and theology. On his return around 1722, he assisted his father at the latter's chapel in London before moving to Tunbridge Wells, Kent, around 1734. There he was minister of the Mount Sion chapel, until 1752.[6]

More generally, we may consider any event B such that $P(B) > 0$ and $k > 1$ events $A_1, \ldots, A_k$ which form a partition of any superset of B (such as the entire sample space S). Then, for any $i = 1,\ldots,k$, it is true that

$$P(A_i \mid B) = \frac{P(A_i B)}{P(B)},$$

where $P(B) = \sum_{j=1}^{n} P(A_j B)$ and $P(A_j B) = P(A_j) P(B \mid A_j)$.

Exercise 1.1 Medical testing

The incidence of a disease in the population is 1%. A medical test for the disease is 90% accurate in the sense that it produces a false reading 10% of the time, both: (a) when the test is applied to a person with the disease; and (b) when the test is applied to a person without the disease.

A person is randomly selected from population and given the test. The test result is positive (i.e. it indicates that the person has the disease).

What is the probability that the person actually has the disease?

Solution to Exercise 1.1

Let A be the event that the person has the disease, and let B be the event that they test positive for the disease. Then:

$P(A) = 0.01$ (the *prior* probability of the person having the disease)

$P(B \mid A) = 0.9$ (the true positive rate, also called the *sensitivity* of the test)

$P(\bar{B} \mid \bar{A}) = 0.9$ (the true negative rate, also called the *specificity* of the test).

So:

$$P(AB) = P(A)P(B \mid A) = 0.01 \times 0.9 = 0.009$$
$$P(\bar{A}B) = P(\bar{A})P(B \mid \bar{A}) = 0.99 \times 0.1 = 0.099.$$

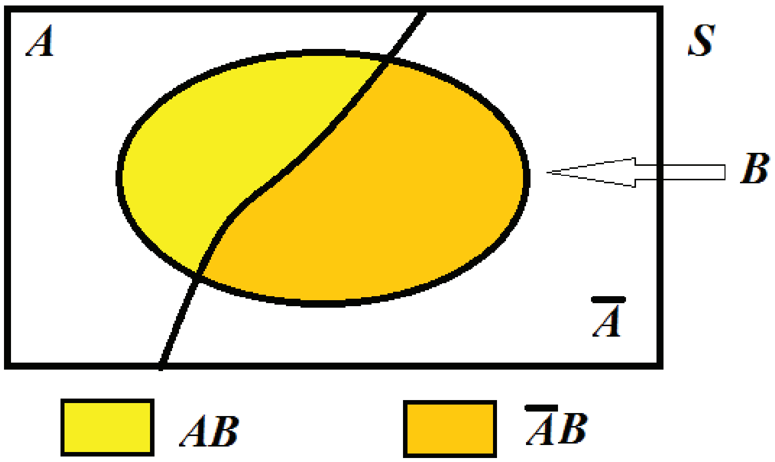
So the unconditional (or prior) probability of the person testing positive is $P(B) = P(AB) + P(\bar{A}B) = 0.009 + 0.099 = 0.108$.

So the required *posterior* probability of the person having the disease is

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{0.009}{0.108} = \frac{1}{12} = 0.08333.$$

Figure 1.2 is a Venn diagram which illustrates how B may be considered as the union of AB and $\bar{A}B$. The required posterior probability of A given B is simply the probability of AB divided by the probability of B.

Figure 1.2 Venn diagram for Exercise 1.1



Discussion

It may seem the posterior probability that the person has the disease (1/12) is rather low, considering the high accuracy of the test (namely $P(B \mid A) = P(\bar{B} \mid \bar{A}) = 0.9$).

This may be explained by considering 1,000 random persons in the population and applying the test to each one. About 10 persons will have the disease, and of these, 9 will test positive. Of the 990 who do not have the disease, 99 will test positive. So the total number of persons testing positive will be 9 + 99 = 108, and the proportion of these 108 who actually have the disease will be 9/108 = 1/12. This heuristic derivation of the answer shows it to be small on account of the large number of false positives (99) amongst the overall number of positives (108).

On the other hand, it may be noted that the posterior probability of the person having the disease is actually very *high* relative to the prior probability of them having the disease ($P(A) = 0.01$). The positive test result has greatly increased the person's chance of having the disease (increased it by more than 700%, since 0.01 + 7.333 × 0.01 = 0.08333).

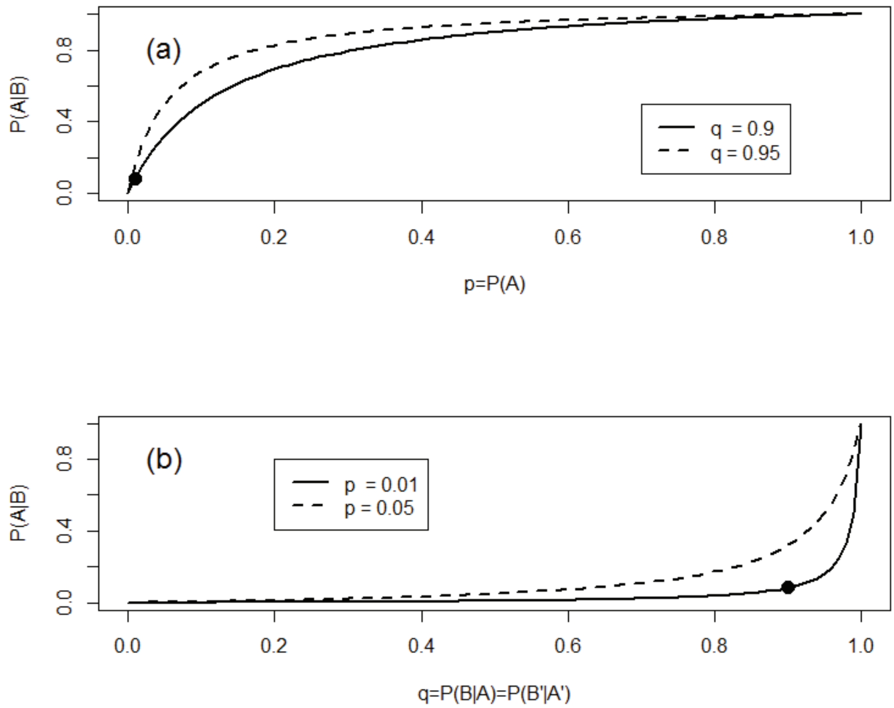
It is instructive to generalise the answer (1/12) as a function of the prevalence (i.e. proportion) of the disease in the population, $p = P(A)$, and the common accuracy rate of the test, $q = P(B \mid A) = P(\bar{B} \mid \bar{A})$.

We find that

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\bar{A})P(B \mid \bar{A})} = \frac{pq}{pq + (1-p)(1-q)}.$$

Figure 1.3 shows the posterior probability of the person having the disease ($P(A \mid B)$) as a function of p with q fixed at 0.9 and 0.95, respectively (subplot (a)), and as a function of q with p fixed at 0.01 and 0.05, respectively (subplot (b)). In each case, the answer (1/12) is represented as a dot corresponding to p = 0.01 and q = 0.9.

Figure 1.3 Posterior probability of disease as functions of *p* and *q*



R Code for Exercise 1.1

```
PAgBfun=function(p=0.01,q=0.9){    p*q   /  (p*q+(1-p)*(1-q))    }
PAgBfun()  # 0.08333333

pvec=seq(0,1,0.01); Pveca=PAgBfun(p=pvec,q=0.9)
          Pveca2=PAgBfun(p=pvec,q=0.95)
qvec=seq(0,1,0.01); Pvecb=PAgBfun(p=0.01,q=qvec)
          Pvecb2=PAgBfun(p=0.05,q=qvec)

X11(w=8,h=7); par(mfrow=c(2,1));

plot(pvec,Pveca,type="l",xlab="p=P(A)",ylab="P(A|B)",lwd=2)
points(0.01,1/12,pch=16,cex=1.5); text(0.05,0.8,"(a)",cex=1.5)
lines(pvec,Pveca2,lty=2,lwd=2)
legend(0.7,0.5,c("q  = 0.9","q = 0.95"),lty=c(1,2),lwd=c(2,2))

plot(qvec,Pvecb,type="l",xlab="q=P(B|A)=P(B'|A')",ylab="P(A|B)",lwd=2)
points(0.9,1/12,pch=16,cex=1.5); text(0.05,0.8,"(b)",cex=1.5)
lines(qvec,Pvecb2,lty=2,lwd=2)
legend(0.2,0.8,c("p  = 0.01","p = 0.05"),lty=c(1,2),lwd=c(2,2))

# Technical note: The graph here was copied from R as ‘bitmap’ and then
# pasted into a Word document, which was then saved as a PDF. If the graph
# is copied from R as ‘metafile’, it appears correct in the Word document,
# but becomes corrupted in the PDF, with axis legends slightly off-centre.
# So, all graphs in this book created in R were copied into Word as ‘bitmap’.
```

Exercise 1.2 Blood types

In a particular population:

10% of persons have Type 1 blood,
and of these, 2% have a particular disease;
30% of persons have Type 2 blood,
and of these, 4% have the disease;
60% of persons have Type 3 blood,
and of these, 3% have the disease.

A person is randomly selected from the population and found to have the disease.

What is the probability that this person has Type 3 blood?

Solution to Exercise 1.2

Let: A = 'The person has Type 1 blood'
B = 'The person has Type 2 blood'
C = 'The person has Type 3 blood'
D = 'The person has the disease'.

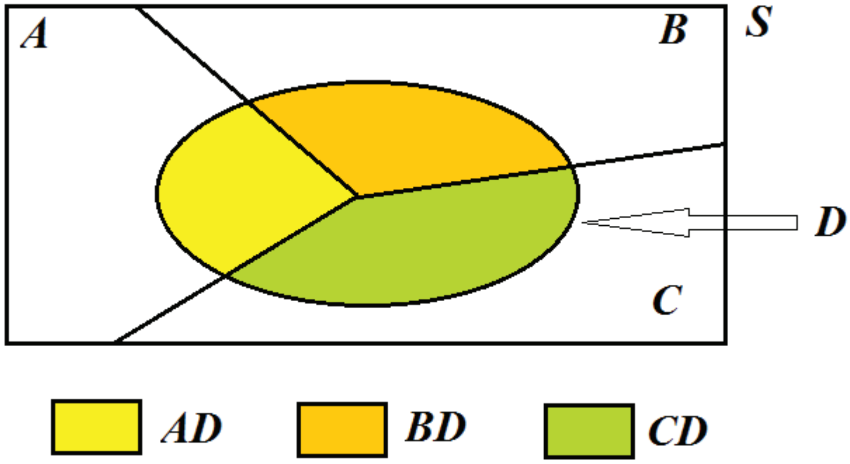
Then: $P(A)=0.1$, $P(D|A)=0.02$
$P(B)=0.3$, $P(D|B)=0.04$
$P(C)=0.6$, $P(D|C)=0.03$.

So:
$$\begin{aligned}P(D) &= P(AD)+P(BD)+P(CD)\\ &= P(A)P(D|A)+P(B)P(D|B)+P(C)P(D|C)\\ &= 0.1\times 0.02+0.3\times 0.04+0.6\times 0.03\\ &= 0.002+0.012+0.018=0.032\,.\end{aligned}$$

Hence: $P(C|D)=\dfrac{P(CD)}{P(D)}=\dfrac{0.018}{0.032}=\dfrac{9}{16}$ = 56.25%.

Figure 1.4 is a Venn diagram showing how D may be considered as the union of AD, BD and CD. The required posterior probability of C given D is simply the probability of CD divided by the probability of D.

Figure 1.4 Venn diagram for Exercise 1.2



1.3 Bayes factors

One way to perform hypothesis testing in the Bayesian framework is via the theory of *Bayes factors*. Suppose that on the basis of an observed event D (standing for *data*) we wish to test a *null hypothesis*

$$H_0 : E_0$$

versus an *alternative hypothesis*

$$H_1 : E_1 ,$$

where E_0 and E_1 are two *events* (which are not necessarily mutually exclusive or even exhaustive of the event space).

Then we calculate:

$\pi_0 = P(E_0)$ = the prior probability of the null hypothesis

$\pi_1 = P(E_1)$ = the prior probability of the alternative hypothesis

$PRO = \pi_0 / \pi_1$ = the prior odds in favour of the null hypothesis

$p_0 = P(E_0 | D)$ = the posterior probability of the null hypothesis

$p_1 = P(E_1 | D)$ = the posterior probability of the alternative hypothesis

$POO = p_0 / p_1$ = the posterior odds in favour of the null hypothesis.

The *Bayes factor* is then defined as $BF = POO / PRO$. This may be interpreted as the factor by which the data have multiplied the odds in favour of the null hypothesis relative to the alternative hypothesis. If $BF > 1$ then the data has *increased* the relative likelihood of the null, and if $BF < 1$ then the data has *decreased* that relative likelihood. The magnitude of BF tells us how much effect the data has had on the relative likelihood.

Note 1: Another way to express the Bayes factor is as

$$BF = \frac{p_0 / p_1}{\pi_0 / \pi_1} = \frac{P(E_0 | D) / P(E_1 | D)}{P(E_0) / P(E_1)} = \frac{P(D)P(E_0 | D) / P(E_0)}{P(D)P(E_1 | D) / P(E_1)} = \frac{P(D | E_0)}{P(D | E_1)}.$$

Thus, the Bayes factor may also be interpreted as the ratio of the likelihood of the data given the null hypothesis to the likelihood of the data given the alternative hypothesis.

Note 2: The idea of a Bayes factor extends to situations where the null and alternative hypotheses are *statistical models* rather than *events*. This idea may be taken up later.

Exercise 1.3 Bayes factor in disease testing

The incidence of a disease in the population is 1%. A medical test for the disease is 90% accurate in the sense that it produces a false reading 10% of the time, both: (a) when the test is applied to a person with the disease; and (b) when the test is applied to a person without the disease.

A person is randomly selected from population and given the test. The test result is positive (i.e. it indicates that the person has the disease).

Calculate the Bayes factor for testing that the person has the disease versus that they do not have the disease.

Solution to Exercise 1.3

Recall in Exercise 1.1, where A = 'Person has disease' and B = 'Person tests positive', the relevant probabilities are $P(A)=0.01$, $P(B|A)=0.9$ and $P(\bar{B}|\bar{A})=0.9$, from which can be deduced that $P(A|B)=1/12$.

We now wish to test $H_0 : A$ vs $H_1 : \bar{A}$. So we calculate:

$\pi_0 = P(A) = 0.01$, $\pi_1 = P(\bar{A}) = 0.99$, $PRO = \pi_0 / \pi_1 = 1/99$,

$p_0 = P(A|B) = 1/12$, $p_1 = P(\bar{A}|B) = 11/12$, $POO = p_0 / p_1 = 1/11$.

Hence the required Bayes factor is $BF = POO/PRO = (1/11)/(1/99) = 9$.

This means the positive test result has multiplied the odds of the person having the disease relative to not having it by a factor of 9 or 900%. Another way to say this is that those odds have increased by 800%.

Note: We could also work out the Bayes factor here as

$$BF = \frac{P(B|A)}{P(B|\bar{A})} = \frac{0.9}{0.1} = 9,$$

namely as the ratio of the probability that the person tests positive given they have the disease to the probability that they test positive given they do not have the disease.

1.4 Bayesian models

Bayes' formula extends naturally to statistical models. A *Bayesian model* is a parametric model in the classical (or frequentist) sense, but with the addition of a *prior probability distribution* for the model parameter, which is treated as a *random variable* rather than an *unknown constant*. The basic components of a Bayesian model may be listed as:

- the *data*, denoted by y
- the *parameter*, denoted by θ
- the *model distribution*, given by a specification of
  $f(y|\theta)$ or $F(y|\theta)$ or the distribution of $(y|\theta)$
- the *prior distribution*, given by a specification of
  $f(\theta)$ or $F(\theta)$ or the distribution of θ.

Here, F is a generic symbol which denotes *cumulative distribution function* (cdf), and f is a generic symbol which denotes *probability density function* (pdf) (when applied to a continuous random variable) or *probability mass function* (pmf) (when applied to a discrete random variable). For simplicity, we will avoid the term pmf and use the term pdf or density for all types of random variable, including the mixed type.

Note 1: A *mixed distribution* is defined by a cdf which exhibits at least one discontinuity (or jump) and is strictly increasing over at least one interval of values.

Note 2: The prior may be specified by writing a statement of the form '$\theta \sim ...$', where the symbol '~' means 'is distributed as', and where '...' denotes the relevant distribution. Likewise, the model for the data may be specified by writing a statement of the form '$(y|\theta) \sim ...$'.

Note 3: At this stage we will not usually distinguish between y as a random variable and y as a value of that random variable; but sometimes we may use Y for the former. Each of y and θ may be a scalar, vector, matrix or array. Also, each component of y and θ may have a discrete distribution, a continuous distribution, or a mixed distribution.

In the first few examples below, we will focus on the simplest case where both y and θ are scalar and discrete.

1.5 The posterior distribution

Bayesian inference requires determination of the *posterior probability distribution* of θ. This task is equivalent to finding the *posterior pdf* of θ, which may be done using the equation

$$f(\theta \mid y) = \frac{f(\theta) f(y \mid \theta)}{f(y)}.$$

Here, $f(y)$ is the *unconditional* (or *prior*) pdf of *y*, as given by

$$f(y) = \int f(y \mid \theta) dF(\theta) = \begin{cases} \int f(\theta) f(y \mid \theta) d\theta & \text{if } \theta \text{ is continuous} \\ \sum_{\theta} f(\theta) f(y \mid \theta) & \text{if } \theta \text{ is discrete.} \end{cases}$$

Note: Here, $\int f(y \mid \theta) dF(\theta)$ is a *Lebesgue-Stieltjes integral*, which may need evaluating by breaking the integral into two parts in the case where θ has a mixed distribution. In the continuous case, think of $dF(\theta)$ as $\frac{dF(\theta)}{d\theta} d\theta = f(\theta) d\theta$.

Exercise 1.4 Loaded dice

Consider six loaded dice with the following properties. Die A has probability 0.1 of coming up 6, each of Dice B and C has probability 0.2 of coming up 6, and each of Dice D, E and F has probability 0.3 of coming up 6.

A die is chosen randomly from the six dice and rolled twice. On both occasions, 6 comes up.

What is the posterior probability distribution of θ, the probability of 6 coming up on the chosen die.

Solution to Exercise 1.4

Let y be the number of times that 6 comes up on the two rolls of the chosen die, and let θ be the probability of 6 coming up on a single roll of that die. Then the Bayesian model is:

$$(y \mid \theta) \sim Bin(2,\theta)$$

$$f(\theta) = \begin{cases} 1/6, & \theta = 0.1 \\ 2/6, & \theta = 0.2 \\ 3/6, & \theta = 0.3. \end{cases}$$

In this case y = 2 and so

$$f(y \mid \theta) = \binom{2}{y} \theta^y (1-\theta)^{2-y} = \binom{2}{2} \theta^2 (1-\theta)^{2-2} = \theta^2 .$$

So $f(y) = \sum_{\theta} f(\theta) f(y \mid \theta) = \frac{1}{6}(0.1)^2 + \frac{2}{6}(0.2)^2 + \frac{3}{6}(0.3)^2 = 0.06.$

So $f(\theta \mid y) = \frac{f(\theta) f(y \mid \theta)}{f(y)} = \begin{cases} (1/6)0.1^2 / 0.06 = 0.02778, & \theta = 0.1 \\ (2/6)0.2^2 / 0.06 = 0.22222, & \theta = 0.2 \\ (3/6)0.3^2 / 0.06 = 0.75, & \theta = 0.3. \end{cases}$

Note: This result means that if the chosen die were to be tossed again a large number of times (say 10,000) then there is a 75% chance that 6 would come up about 30% of the time, a 22.2% chance that 6 would come up about 20% of the time, and a 2.8% chance that 6 would come up about 10% of the time.

1.6 The proportionality formula

Observe that $f(y)$ is a constant with respect to θ in the Bayesian equation

$$f(\theta \mid y) = f(\theta) f(y \mid \theta) / f(y) ,$$

which means that we may also write the equation as

$$f(\theta \mid y) = \frac{f(\theta) f(y \mid \theta)}{k} ,$$

or as

$$f(\theta \mid y) = cf(\theta) f(y \mid \theta) ,$$

where $k = f(y)$ and $c = 1/k$.

We may also write

$$f(\theta \mid y) \propto f(\theta) f(y \mid \theta) ,$$

where $\propto$ is the proportionality sign.

Equivalently, we may write

$$f(\theta \mid y) \overset{\theta}{\propto} f(\theta) f(y \mid \theta)$$

to emphasise that the proportionality is specifically with respect to θ.

Another way to express the last equation is

$$f(\theta \mid y) \propto f(\theta) \times L(\theta \mid y),$$

where $L(\theta \mid y)$ is the *likelihood function* (defined as the model density $f(y \mid \theta)$ multiplied by any constant with respect to θ, and viewed as a function of θ rather than of y).

The last equation may also be stated in words as:

The posterior is proportional to the prior times the likelihood.

These observations indicate a shortcut method for determining the required posterior distribution which obviates the need for calculating $f(y)$ (which may be difficult).

This method is to multiply the prior density (or the kernel of that density) by the likelihood function and try to identify the resulting function of θ as the density of a well-known or common distribution.

Once the posterior distribution has been identified, $f(y)$ may then be obtained easily as the associated normalising constant.

Exercise 1.5 Loaded dice with solution via the proportionality formula

As in Exercise 1.4, suppose that Die A has probability 0.1 of coming up 6, each of Dice B and C has probability 0.2 of coming up 6, and each of Dice D, E and F has probability 0.3 of coming up 6.

A die is chosen randomly from the six dice and rolled twice. On both occasions, 6 comes up.

Using the proportionality formula, find the posterior probability distribution of θ, the probability of 6 coming up on the chosen die.

Solution to Exercise 1.5

With y denoting the number of times 6 comes up, the Bayesian model may be written:

$$f(y\,|\,\theta)=\binom{2}{y}\theta^{y}(1-\theta)^{2-y},\ y=0,1,2$$

$$f(\theta)=10\theta/6,\ \theta=0.1,0.2,0.3\,.$$

Note: $10\theta/6$ = 1/6, 2/6 and 3/6 for θ = 0.1, 0.2 and 0.3, respectively.

Hence $f(\theta\,|\,y)\propto f(\theta)f(y\,|\,\theta)$

$$=\frac{10\theta}{6}\times\binom{2}{y}\theta^{y}(1-\theta)^{2-y}$$

$$\propto\theta\times\theta^{2}\qquad\text{since } y=2.$$

Thus $f(\theta\,|\,y)\propto\theta^{3}=\begin{cases}0.1^{3}=1/1000, & \theta=0.1\\ 0.2^{3}=8/1000, & \theta=0.2\\ 0.3^{3}=27/1000, & \theta=0.3\end{cases}\propto\begin{cases}1, & \theta=0.1\\ 8, & \theta=0.2\\ 27, & \theta=0.3.\end{cases}$

Now, $1+8+27=36$, and so $f(\theta\,|\,y)=\begin{cases}1^{3}/36=0.02778, & \theta=0.1\\ 2^{3}/36=0.22222, & \theta=0.2\\ 3^{3}/36=0.75, & \theta=0.3,\end{cases}$

which is the same result as obtained earlier in Exercise 1.4.

Exercise 1.6 Buses

You are visiting a town with buses whose licence plates show their numbers consecutively from 1 up to however many there are. In your mind the number of buses could be anything from one to five, with all possibilities equally likely.

Whilst touring the town you first happen to see Bus 3.

Assuming that at any point in time you are equally likely to see any of the buses in the town, how likely is it that the town has at least four buses?

Solution to Exercise 1.6

Let θ be the number of buses in the town and let y be the number of the bus that you happen to first see. Then an appropriate Bayesian model is:

$$f(y|\theta) = 1/\theta,\ y = 1,...,\theta$$
$$f(\theta) = 1/5,\ \theta = 1,...,5 \quad \text{(prior).}$$

Note: We could also write this model as:

$$(y|\theta) \sim DU(1,...,\theta)$$
$$\theta \sim DU(1,...,5)\,,$$

where DU denotes the *discrete uniform distribution*. (See Appendix B.9 for details regarding this distribution. Appendix B also provides details regarding some other important distributions that feature in this book.)

So the posterior density of θ is

$$f(\theta|y) \propto f(\theta)f(y|\theta)$$
$$\propto 1 \times 1/\theta\,,\quad \theta = y,...,5\,.$$

Noting that $y = 3$, we have that

$$f(\theta|y) \propto \begin{cases} 1/3, \theta = 3 \\ 1/4, \theta = 4 \\ 1/5, \theta = 5. \end{cases}$$

Now, $1/3 + 1/4 + 1/5 = (20+15+12)/60 = 47/60$, and so

$$f(\theta|y) = \begin{cases} \dfrac{1/3}{47/60} = \dfrac{20}{47}, \theta = 3 \\ \dfrac{1/4}{47/60} = \dfrac{15}{47}, \theta = 4 \\ \dfrac{1/5}{47/60} = \dfrac{12}{47}, \theta = 5. \end{cases}$$

So the posterior probability that the town has at least four buses is

$$P(\theta \geq 4|y) = \sum_{\theta:\theta \geq 4} f(\theta|y) = f(\theta = 4|y) + f(\theta = 5|y)$$
$$= 1 - f(\theta = 3|y) = 1 - \frac{20}{47} = \frac{27}{47} = 0.5745.$$

Discussion

This exercise is a variant of the famous 'tramcar problem' considered by Harold Jeffreys in his book *Theory of Probability* and previously suggested to him by M.H.A. Newman (see Jeffreys, 1961, page 238). Suppose that before entering the town you had *absolutely no idea* about the number of buses θ. Then, according to Jeffreys' logic, a prior which may be considered as suitably uninformative (or noninformative) in this situation is given by $f(\theta) \propto 1/\theta$, $\theta = 1, 2, 3, \ldots$.

Now, this prior density is problematic because it is *improper* (since $\sum_{\theta=1}^{\infty} 1/\theta = \infty$). However, it leads to a *proper* posterior density given by

$$f(\theta \mid y) = \frac{1}{c\theta^2}, \ \theta = 3, 4, 5, \ldots,$$

where $c = \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \ldots = \frac{\pi^2}{6} - \left(\frac{1}{1^2} + \frac{1}{2^2}\right)$ = 0.394934.

So, under this alternative prior, the probability of there being at least four buses in the town (given that you have seen Bus 3) works out as

$$P(\theta \geq 4 \mid y) = 1 - P(\theta = 3 \mid y) = 1 - \frac{1}{9c} = 0.7187.$$

The logic which Jeffreys used to come up with the prior $f(\theta) \propto 1/\theta$ in relation to the tramcar problem will be discussed further in Chapter 2.

R Code for Exercise 1.6

```
options(digits=6); c=(1/6)*(pi^2)-5/4; c # 0.394934
1- (1/3^2)/c # 0.718659
```

Exercise 1.7 Balls in a box

In each of nine indistinguishable boxes there are nine balls, the *i*th box having *i* red balls and $9 - i$ white balls (*i* = 1,…,9).

One box is selected randomly from the nine, and then three balls are chosen randomly from the selected box (without replacement and without looking at the remaining balls in the box).

Exactly two of the three chosen balls are red. Find the probability that the selected box has at least four red balls remaining in it.

Solution to Exercise 1.7

Let: N = the number of balls in each box (9)
n = the number of balls chosen from the selected box (3)
θ = the number of red balls initially in the selected box (1,2,…,8 or 9)
y = the number of red balls amongst the n chosen balls (2).

Then an appropriate Bayesian model is:

$(y\,|\,\theta) \sim Hyp(N,\theta,n)$ (Hypergeometric with parameters N, θ and n, and having mean $n\theta/N$)

$\theta \sim DU(1,...,N)$ (discrete uniform over the integers 1,2,…,N).

For this model, the posterior density of θ is

$$f(\theta\,|\,y) \propto f(\theta)f(y\,|\,\theta) = \frac{1}{N}\times\binom{\theta}{y}\binom{N-\theta}{n-y}\bigg/\binom{N}{n}$$

$$\propto \frac{\theta!(N-\theta)!}{(\theta-y)!(N-\theta-(n-y))!}, \quad \theta = y,...,N-(n-y).$$

In our case,

$$f(\theta\,|\,y) \propto \frac{\theta!(9-\theta)!}{(\theta-2)!(9-\theta-(3-2))!}, \; \theta = 2,...,9-(3-2),$$

or more simply,

$$f(\theta\,|\,y) \propto \theta(\theta-1)(9-\theta), \quad \theta = 2,...,8.$$

Thus $$f(\theta\,|\,y) \propto \begin{Bmatrix} 14, \theta=2 \\ 36, \theta=3 \\ 60, \theta=4 \\ 80, \theta=5 \\ 90, \theta=6 \\ 84, \theta=7 \\ 56, \theta=8 \end{Bmatrix} \equiv k(\theta),$$

where

$$c \equiv \sum_{\theta=1}^{8} k(\theta) = 14 + 36 + \ldots + 56 = 420.$$

So $$f(\theta \mid y) = \frac{k(\theta)}{c} = \begin{cases} 14/420 = 0.03333, \theta = 2 \\ 36/420 = 0.08571, \theta = 3 \\ 60/420 = 0.14286, \theta = 4 \\ 80/420 = 0.19048, \theta = 5 \\ 90/420 = 0.21429, \theta = 6 \\ 84/420 = 0.20000, \theta = 7 \\ 56/420 = 0.13333, \theta = 8. \end{cases}$$

The probability that the selected box has at least four red balls remaining is the posterior probability that θ (the number of red balls initially in the box) is at least 6 (since two red balls have already been taken out of the box). So the required probability is

$$P(\theta \geq 6 \mid y) = \frac{90+84+56}{420} = \frac{23}{42} = 0.5476.$$

R Code for Exercise 1.7

```
tv=2:8; kv=tv*(tv-1)*(9-tv); c=sum(kv); c # 420
options(digits=4);  cbind(tv,kv,kv/c,cumsum(kv/c))
# [1,]  2   14   0.03333   0.03333
# [2,]  3   36   0.08571   0.11905
# [3,]  4   60   0.14286   0.26190
# [4,]  5   80   0.19048   0.45238
# [5,]  6   90   0.21429   0.66667
# [6,]  7   84   0.20000   0.86667
# [7,]  8   56   0.13333   1.00000

23/42 # 0.5476
1-0.45238 # 0.5476  (alternative calculation of the required probability)
sum((kv/c)[tv>=6]) #  0.5476
                    # (yet another calculation of the required probability)
```

1.7 Continuous parameters

The examples above have all featured a target parameter which is *discrete*. The following example illustrates Bayesian inference involving a *continuous* parameter. This case presents no new problems, except that the prior and posterior densities of the parameter may no longer be interpreted directly as probabilities.

Exercise 1.8 The *binomial-beta model* (or *beta-binomial model*)

Consider the following Bayesian model:

$$(y \mid \theta) \sim Binomial(n, \theta)$$
$$\theta \sim Beta(\alpha, \beta) \quad \text{(prior).}$$

Find the posterior distribution of θ.

Solution to Exercise 1.8

The posterior density is

$$\begin{aligned}
f(\theta \mid y) &\propto f(\theta) f(y \mid \theta) \\
&= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \times \binom{n}{y} \theta^{y}(1-\theta)^{n-y} \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^{y}(1-\theta)^{n-y} \quad \text{(ignoring constants which do not depend on } \theta\text{)} \\
&= \theta^{(\alpha+y)-1}(1-\theta)^{(\beta+n-y)-1}, \; 0<\theta<1.
\end{aligned}$$

This is the kernel of the beta density with parameters $\alpha + y$ and $\beta + n - y$. It follows that the posterior distribution of θ is given by

$$(\theta \mid y) \sim Beta(\alpha + y, \beta + n - y),$$

and the posterior density of θ is (exactly)

$$f(\theta \mid y) = \frac{\theta^{(\alpha+y)-1}(1-\theta)^{(\beta+n-y)-1}}{B(\alpha+y, \beta+n-y)}, \; 0<\theta<1.$$

For example, suppose that $\alpha = \beta = 1$, that is, $\theta \sim Beta(1,1)$.

Then the prior density is $f(\theta) = \dfrac{\theta^{1-1}(1-\theta)^{1-1}}{B(1,1)} = 1, \; 0<\theta<1$.

Thus the prior may also be expressed by writing $\theta \sim U(0,1)$.

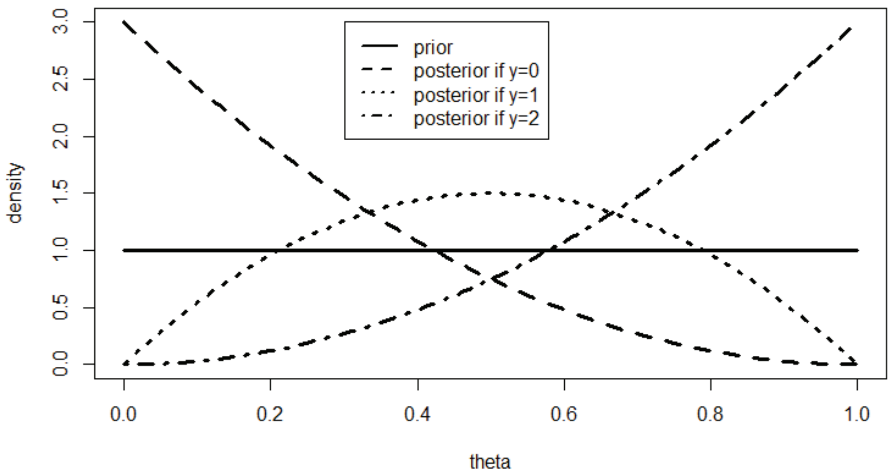
Also, suppose that $n = 2$. Then there are three possible values of y, namely 0, 1 and 2, and these lead to the following three posteriors, respectively:

$$(\theta \mid y) \sim Beta(1+0, 1+2-0) = Beta(1,3)$$
$$(\theta \mid y) \sim Beta(1+1, 1+2-1) = Beta(2,2)$$
$$(\theta \mid y) \sim Beta(1+2, 1+2-2) = Beta(3,1).$$

These three posteriors and the prior are illustrated in Figure 1.5.

Note: The prior here may be considered uninformative because it is 'flat' over the entire range of possible values for θ, namely 0 to 1. This prior was originally used by Thomas Bayes and is often called the *Bayes prior*. However, other uninformative priors have been proposed for the binomial parameter θ. These will be discussed later, in Chapter 2.

Figure 1.5 The prior and three posteriors in Exercise 1.8



R Code for Exercise 1.8

```
X11(w=8,h=5); par(mfrow=c(1,1));

plot(c(0,1),c(0,3),type="n",xlab="theta",ylab="density")

lines(c(0,1),c(1,1),lty=1,lwd=3);  tv=seq(0,1,0.01)
lines(tv,3*(1-tv)^2,lty=2,lwd=3)
lines(tv,3*2*tv*(1-tv),lty=3,lwd=3)
lines(tv,3*tv^2,lty=4,lwd=3)

legend(0.3,3,c("prior","posterior if y=0","posterior if y=1","posterior if y=2"),
   lty=c(1,2,3,4),lwd=rep(2,4))
```

1.8 Finite and infinite population inference

In the last example (Exercise 1.8), with the model:

$$(y\,|\,\theta) \sim Binomial(n,\theta)$$

$$\theta \sim Beta(\alpha,\beta)\,,$$

the quantity of interest θ is the probability of success on a single Bernoulli trial.

This quantity may be thought of as the average of a hypothetically *infinite* number of Bernoulli trials. For that reason we may refer to derivation of the posterior distribution,

$$(\theta\,|\,y) \sim Beta(\alpha+y,\beta+n-y)\,,$$

as *infinite population inference*.

In contrast, for the 'buses' example further above (Exercise 1.6), which involves the model:

$$f(y\,|\,\theta)=1/\theta,\; y=1,\ldots,\theta$$

$$f(\theta)=1/5,\, \theta=1,\ldots,5\,,$$

the quantity of interest θ represents the number of buses in a population of buses, which of course is *finite*.

Therefore derivation of the posterior,

$$f(\theta\,|\,y)=\begin{cases}20/47, \theta=3\\ 15/47, \theta=4\\ 12/47, \theta=5,\end{cases}$$

may be termed *finite population inference*.

Another example of finite population inference is the 'balls in a box' example (Exercise 1.7), where the model is:

$$(y\,|\,\theta) \sim Hyp(N,\theta,n)$$

$$\theta \sim DU(1,\ldots,N)\,,$$

and where the quantity of interest θ is the number of red balls initially in the selected box (1,2,…,8 or 9).

And another example of *infinite* population inference is the 'loaded dice' example (Exercises 1.4 and 1.5), where the model is:

$$f(y\,|\,\theta)=\binom{2}{y}\theta^y(1-\theta)^{2-y},\; y=0,1,2$$

$$f(\theta)=10\theta/6,\; \theta=0.1,0.2,0.3,$$

and where the quantity of interest θ is the probability of 6 coming up on a single roll of the chosen die (i.e. the average number of 6s that come up on a hypothetically infinite number of rolls of that particular die).

Generally, finite population inference may also be thought of in terms of *prediction* (e.g. in the 'buses' example, we are *predicting* the total number of buses in the town). For that reason, finite population inference may also be referred to as *predictive inference*. Yet another term for finite population inference is *descriptive inference*. In contrast, infinite population inference may also be called *analytic inference*. More will be said on finite population/predictive/descriptive inference in later chapters of the course.

1.9 Continuous data

So far, all the Bayesian models considered have featured data which is modelled using a *discrete* distribution. (Some of these models have a discrete parameter and some have a continuous parameter.) The following is an example with data that follows a *continuous* probability distribution. (This example also has a continuous parameter.)

Exercise 1.9 The *exponential-exponential model*

Suppose θ has the standard exponential distribution, and the conditional distribution of y given θ is exponential with mean $1/\theta$. Find the posterior density of θ given y.

Solution to Exercise 1.9

The Bayesian model here is: $f(y\,|\,\theta) = \theta e^{-\theta y},\, y > 0$

$$f(\theta) = e^{-\theta},\, \theta > 0\,.$$

So $f(\theta\,|\,y) \propto f(\theta) f(y\,|\,\theta) \propto e^{-\theta} \times \theta e^{-\theta y} = \theta^{2-1} e^{-\theta(y+1)},\, y > 0.$

This is the kernel of a gamma distribution with parameters 2 and $y + 1$, as per the definitions in Appendix B.2. Thus we may write

$$(\theta\,|\,y) \sim Gamma(2, y+1),$$

from which it follows that the posterior density of θ is

$$f(\theta\,|\,y) = \frac{(y+1)^2 \theta^{2-1} e^{-\theta(y+1)}}{\Gamma(2)},\, \theta > 0\,.$$

Exercise 1.10 The *uniform-uniform model*

Consider the Bayesian model given by:

$$(y \mid \theta) \sim U(0,\theta)$$
$$\theta \sim U(0,1)\,.$$

Find the posterior density of θ given *y*.

Solution to Exercise 1.10

Noting that $0 < y < \theta < 1$, we see that the posterior density is

$$f(\theta \mid y) = \frac{f(\theta)\, f(y \mid \theta)}{f(y)} = \frac{1\times(1/\theta)}{\int_y^1 1\times(1/\theta)d\theta}$$
$$= \frac{1/\theta}{\log 1 - \log y} = \frac{-1}{\theta \log y},\ y < \theta < 1\,.$$

Note: This is a 'non-standard' density and strictly decreasing. To give a physical example, a stick of length 1 metre is cut at a point randomly located along its length. The part to the right of the cut is discarded and then another cut is made randomly along the stick which remains. Then the part to the right of that second cut is likewise discarded. The length of the stick remaining after the first cut is a random variable with density as given above, with *y* being the length of the finally remaining stick.

1.10 Conjugacy

When the prior and posterior distributions are members of the same class of distributions, we say that they form a *conjugate pair*, or that the prior is *conjugate*. For example, consider the binomial-beta model:

$$(y \mid \theta) \sim Binomial(n,\theta)$$
$$\theta \sim Beta(\alpha,\beta) \qquad \text{(prior)}$$
$$\Rightarrow \quad (\theta \mid y) \sim Beta(\alpha + y, \beta + n - y) \qquad \text{(posterior).}$$

Since both prior and posterior are beta, the prior is conjugate.

Likewise, consider the exponential-exponential model:

$$f(y \mid \theta) = \theta e^{-\theta y},\ y > 0$$
$$f(\theta) = e^{-\theta},\ \theta > 0 \qquad \text{(i.e. } \theta \sim Gamma(1,1)\text{)} \qquad \text{(prior)}$$
$$\Rightarrow \quad (\theta \mid y) \sim Gamma(2, y+1) \qquad \text{(posterior).}$$

Since both prior and posterior are gamma, the prior is conjugate.

On the other hand, consider the model in the buses example:

$$(y \mid \theta) \sim DU(1,\ldots,\theta)$$

$$\theta \sim DU(1,\ldots,5) \qquad \text{(prior)}$$

$$\Rightarrow \quad f(\theta \mid y=3) = \begin{cases} 20/47, \theta = 3 \\ 15/47, \theta = 4 \\ 12/47, \theta = 5 \end{cases} \qquad \text{(posterior).}$$

The prior is discrete uniform but the posterior is not. So in this case the prior is not conjugate.

Specifying a Bayesian model using a conjugate prior is generally desirable because it can simplify the calculations required.

1.11 Bayesian point estimation

Once the posterior distribution or density $f(\theta \mid y)$ has been obtained, Bayesian point estimates of the model parameter θ can be calculated. The three most commonly used point estimates are as follows.

- The *posterior mean* of θ is

$$E(\theta \mid y) = \int \theta dF(\theta \mid y) = \begin{cases} \int \theta f(\theta \mid y) d\theta & \text{if } \theta \text{ is continuous} \\ \sum_{\theta} \theta f(\theta \mid y) & \text{if } \theta \text{ is discrete.} \end{cases}$$

- The *posterior mode* of θ is

$Mode(\theta \mid y)$ = any value $m \in \Re$ which satisfies

$$f(\theta = m \mid x) = \max_{\theta} f(\theta \mid x)$$

$$\text{or } \lim_{\theta \to m} f(\theta \mid x) = \sup f(\theta \mid x),$$

or the set of all such values.

- The *posterior median* of θ is

$Median(\theta \mid y)$ = any value m of θ such that

$$P(\theta \leq m \mid y) \geq 1/2$$

$$\text{and } P(\theta \geq m \mid y) \geq 1/2,$$

or the set of all such values.

Note 1: In some cases, the posterior mean does not exist or it is equal to infinity or minus infinity.

Note 2: Typically, the posterior mode and posterior median are unique. The above definitions are given for completeness.

Note 3: The integral $\int \theta dF(\theta | y)$ is a Lebesgue-Stieltje's integral. This may need to be evaluated as the sum of two separate parts in the case where θ has a mixed distribution. In the continuous case, it is useful to think of $dF(\theta | y)$ as $\frac{dF(\theta | y)}{d\theta} d\theta = f(\theta | y) d\theta$.

Note 4: The above three Bayesian point estimates may be interpreted in an intuitive manner. For example, θ's posterior mode is the value of θ which is 'made most likely by the data'. They may also be understood in the context of *Bayesian decision theory* (discussed later).

1.12 Bayesian interval estimation

There are many ways to construct a Bayesian interval estimate, but the two most common ways are defined as follows. The $1-\alpha$ (or $100(1-\alpha)\%$) *highest posterior density region* (HPDR) for θ is the smallest set *S* such that:

$$P(\theta \in S | y) \geq 1-\alpha$$

and $f(\theta_1 | y) \geq f(\theta_2 | y)$ if $\theta_1 \in S$ and $\theta_2 \notin S$.

Figure 1.6 illustrates the idea of the HPDR. In the very common situation where θ is scalar, continuous and has a posterior density which is unimodal with no local modes (i.e. has the form of a single 'mound'), the $1-\alpha$ HPDR takes on the form of a single interval defined by two points at which the posterior density has the same value. When the HPDR is a single interval, it is the shortest possible single interval over which the area under the posterior density is $1-\alpha$.

The $1-\alpha$ *central posterior density region* (CPDR) for a scalar parameter θ may be defined as the shortest single interval [*a*,*b*] such that:

$$P(\theta < a | y) \leq \alpha / 2$$

and $P(\theta > b | y) \leq \alpha / 2$.

Figure 1.6 An 80% HPDR

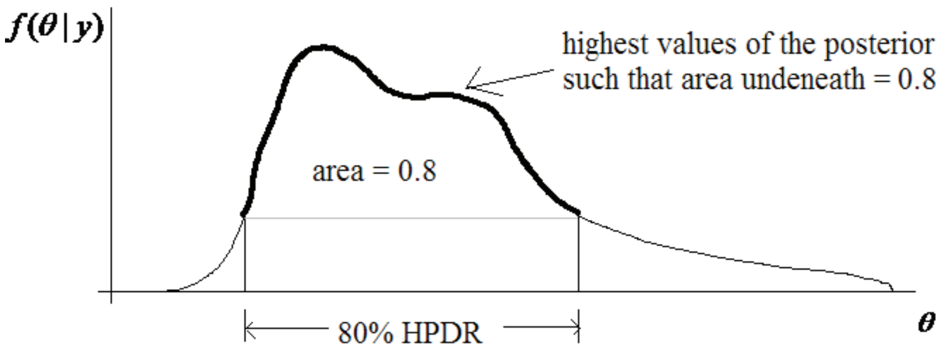


Figure 1.7 illustrates the idea of the CPDR. One drawback of the CPDR is that it is only defined for a *scalar* parameter. Another drawback is that some values *inside* the CPDR may be less likely *a posteriori* than some values *outside* it (which is not the case with the HPDR). For example, in Figure 1.7, a value *just below the upper bound* of the 80% CPDR has a smaller posterior density than a value *just below the lower bound* of that CPDR. However, CPDRs are typically easier to calculate than HPDRs.

In the common case of a continuous parameter with a posterior density in the form of a single 'mound' which is furthermore symmetric, the CPDR and HPDR are identical.

Note 1: The $1-\alpha$ CPDR for θ may alternatively be defined as the shortest single *open* interval (a,b) such that:

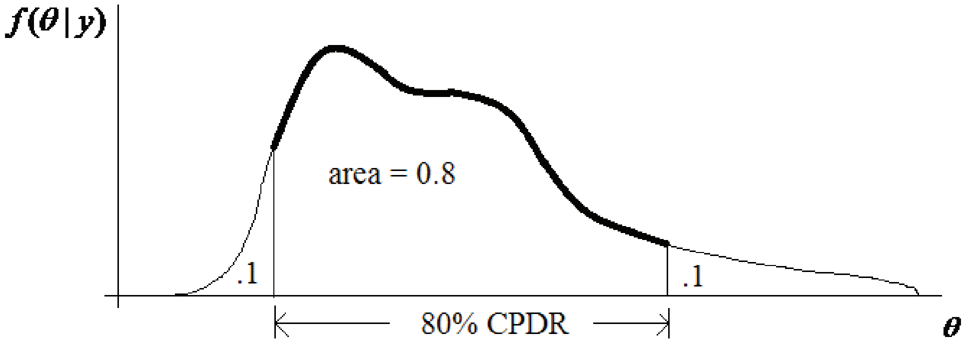
$$P(\theta \le a \mid y) \le \alpha/2$$

and $$P(\theta \ge b \mid y) \le \alpha/2 .$$

Other variations are possible (of the form $[a,b)$ and $(a,b]$); but when the parameter of interest θ is continuous these definitions are all equivalent. Yet another definition of the $1-\alpha$ CPDR is any of the CPDRs as defined above but with all *a posteriori* impossible values of θ excluded.

Note 2: As regards terminology, whenever the HPDR is a single interval, it may also be called the *highest posterior density interval* (HPDI). Likewise, the CPDR, which is always a single interval, may also be called the *central posterior density interval* (CPDI).

Figure 1.7 An 80% CPDR



Exercise 1.11 A bent coin

We have a bent coin, for which θ, the probability of heads coming up, is unknown. Our prior beliefs regarding θ may be described by a standard uniform distribution. Thus no value of θ is deemed more or less likely than any other.

We toss the coin *n* = 5 times (independently), and heads come up every time.

Find the posterior mean, mode and median of θ. Also find the 80% HPDR and CPDR for θ .

Solution to Exercise 1.11

Recall the binomial-beta model:

$$(y\,|\,\theta) \sim Binomial(n,\theta)$$

$$\theta \sim Beta(\alpha,\beta)\,,$$

for which $(\theta\,|\,y) \sim Beta(\alpha+y,\beta+n-y)$.

We now apply this result with $n = y = 5$ and $\alpha=\beta=1$ (corresponding to $\theta \sim U(0,1)$), and find that:

$$(\theta\,|\,y) \sim Beta(1+5,5-5+1) = Beta(6,1)$$

$$f(\theta\,|\,y) = \frac{\theta^{6-1}(1-\theta)^{1-1}}{B(6,1)} = 6\theta^5, \quad 0<\theta<1$$

$$F(\theta\,|\,y) = \int_0^\theta 6t^5 dt = \theta^6, \quad 0<\theta<1.$$

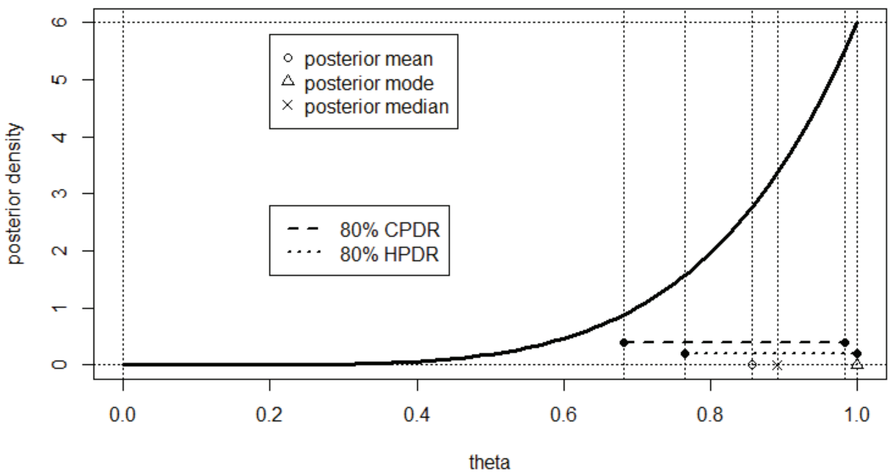
Therefore: $E(\theta \mid y) = \dfrac{6}{6+1} = \dfrac{6}{7} = 0.8571$

$$Mode(\theta \mid y) = \frac{6-1}{(6-1)+(1-1)} = 1$$

$$Median(\theta \mid y) = \text{solution in } \theta \text{ of } F(\theta \mid y) = 1/2, \text{ i.e. } \theta^6 = 0.5$$
$$= (0.5)^{1/6} = 0.8909.$$

Also, the 80% HPDR is $(0.2^{1/6}, 1) =$ (0.7647,1) (since $f(\theta \mid y)$ is strictly increasing), and the 80% CPDR is $(0.1^{1/6}, 0.9^{1/6}) =$ (0.6813,0.9826). The three point estimate and two interval estimates just derived are shown in Figure 1.8.

Figure 1.8 Inference in Exercise 1.11



R Code for Exercise 1.11

```
options(digits=4); postmean=6/7; postmode=1; postmedian=0.5^(1/6)
c(postmean,postmode,postmedian) # 0.8571 1.0000 0.8909
hpdr=c(0.2^(1/6),1); cpdr=c(0.1,0.9)^(1/6)
c(hpdr,cpdr) # 0.7647 1.0000 0.6813 0.9826

X11(w=8,h=5); par(mfrow=c(1,1)); tv=seq(0,1,0.01); fv=dbeta(tv,6,1)
plot(tv,fv,type="l",lwd=3,xlab="theta",ylab="posterior density")
points(c(postmean,postmode,postmedian),c(0,0,0),pch=c(1,2,4))
points(hpdr,rep(0.2,2),pch=16); lines(hpdr,rep(0.2,2),lty=3,lwd=2)
```

```
points(cpdr,rep(0.4,2),pch=16); lines(cpdr,rep(0.4,2),lty=2,lwd=2)
abline(v=c(postmean,postmode,postmedian),lty=3)
abline(v=c(0,hpdr,cpdr),lty=3); abline(h=c(0,6),lty=3)
legend(0.2,5.8,c("posterior mean","posterior mode",
           "posterior median"),pch=c(1,2,4))
legend(0.2,2.8,c("80% CPDR","80% HPDR"),lty=c(2,3),lwd=c(2,2))
```

Exercise 1.12 HPDR and CPDR for a discrete parameter

Consider the posterior distribution from Exercise 1.7 (Balls in a box):

$$f(\theta \mid y) = \begin{cases} 14/420 = 0.03333, \theta = 2 \\ 36/420 = 0.08571, \theta = 3 \\ 60/420 = 0.14286, \theta = 4 \\ 80/420 = 0.19048, \theta = 5 \\ 90/420 = 0.21429, \theta = 6 \\ 84/420 = 0.20000, \theta = 7 \\ 56/420 = 0.13333, \theta = 8. \end{cases}$$

Find the 90% HPDR and 90% CPDR for θ. Also find the 50% HPDR and 50% CPDR for θ. For each region, calculate the associated exact coverage probability.

Solution to Exercise 1.12

The 90% HPDR is the set {3,4,5,6,7,8};
this has exact coverage 1 − 14/420 = 0.9667.

The 90% CPDR is the closed interval [3, 8];
this likewise has exact coverage 0.9667.

The 50% HPDR is {5,6,7};
this has exact coverage (80 + 90 + 84)/420 = 0.6047.

The 50% CPDR is [4, 7];
this has exact coverage (60 + 80 + 90 + 84)/420 = 0.7476.

Note: The lower bound of the 50% CPDR cannot be equal to 5. This is because $P(\theta < 5 \mid y) = (14 + 36 + 60)/420$ = 0.2619, which is not less than or equal to $\alpha / 2 = 0.25$, as required by the definition of CPDR.

Exercise 1.13 Illustration of the definition of HPDR

Suppose that the posterior probabilities of a parameter θ given data y are exactly 10%, 40% and 50% for values 1, 2 and 3, respectively. Find S, the 40% HPDR for θ.

Solution to Exercise 1.13

The smallest set S such that $P(\theta \in S \mid y) \geq 0.4$ is $\{2\}$ or $\{3\}$. With the additional requirement that $f(\theta_1 \mid y) \geq f(\theta_2 \mid y)$ if $\theta_1 \in S$ and $\theta_2 \notin S$, we see that $S = \{3\}$ (only). That is, the 40% HPDR is the singleton set $\{3\}$.

1.13 Inference on functions of the model parameter

So far we have examined Bayesian models with a single parameter θ and described how to perform posterior inference on that parameter. Sometimes there may also be interest in some *function* of the model parameter, denoted by (say)

$$\psi = g(\theta).$$

Then the posterior density of ψ can be derived using distribution theory, for example by applying the transformation rule,

$$f(\psi \mid y) = f(\theta \mid y)\left|\frac{d\theta}{d\psi}\right|,$$

in cases where $\psi = g(\theta)$ is strictly increasing or strictly decreasing.

Point and interval estimates of ψ can then be calculated in the usual way, using $f(\psi \mid y)$. For example, the posterior mean of ψ equals

$$E(\psi \mid y) = \int \psi\, f(\psi \mid y) d\psi .$$

Sometimes it is more practical to calculate point and interval estimates another way, without first deriving $f(\psi \mid y)$.

For example, another expression for the posterior mean is

$$E(\psi \mid y) = E(g(\theta) \mid y) = \int g(\theta) f(\theta \mid y) d\theta .$$

Also, the posterior median of ψ, call this *M*, can typically be obtained by simply calculating

$$M = g(m),$$

where *m* is the posterior median of θ.

Note: To see why this works, we write

$$P(\psi < M \mid y) = P(g(\theta) < M \mid y)$$
$$= P(g(\theta) < g(m) \mid y) = P(\theta < m \mid y) = 1/2.$$

Exercise 1.14 Estimation of an exponential mean

Suppose that θ has the standard exponential distribution, and *y* given θ is exponential with mean $1/\theta$. Find the posterior density and posterior mean of the model mean, $\psi = E(y \mid \theta) = 1/\theta$, given the data *y*.

Solution to Exercise 1.14

Recall that the Bayesian model

$$f(y \mid \theta) = \theta e^{-\theta y},\ y > 0$$
$$f(\theta) = e^{-\theta},\ \theta > 0$$

implies the posterior $(\theta \mid y) \sim Gamma(2, y+1)$.

So, by definition, $(\psi \mid y) \sim InverseGamma(2, y+1)$,

with density $f(\psi \mid y) = \dfrac{(y+1)^2 \psi^{-(2+1)} e^{-(y+1)/\psi}}{\Gamma(2)} = \dfrac{(y+1)^2}{\psi^3 e^{(y+1)/\psi}},\ \psi > 0,$

and mean $E(\psi \mid y) = \dfrac{y+1}{2-1} = y+1$.

Note: This mean could also be obtained as follows:

$$E(\psi \mid y) = E\left(\frac{1}{\theta}\middle| y\right) = \int_0^\infty \frac{1}{\theta} f(\theta \mid y) d\theta$$
$$= \int_0^\infty \frac{1}{\theta} \times \frac{(y+1)^2 \theta^{2-1} e^{-\theta(y+1)}}{\Gamma(2)} d\theta$$
$$= \frac{\Gamma(1)(y+1)^2}{\Gamma(2)(y+1)^1} \int_0^\infty \frac{1}{\theta} \times \frac{(y+1)^1 \theta^{1-1} e^{-\theta(y+1)}}{\Gamma(1)} d\theta$$
$$= y+1 \quad \text{(using the fact that the last integral equals 1).}$$

Exercise 1.15 Inference on a function of the binomial parameter

Recall the binomial-beta model given by:

$$(y \mid \theta) \sim Binomial(n, \theta)$$
$$\theta \sim Beta(\alpha, \beta),$$

for which $(\theta \mid y) \sim Beta(\alpha + y, \beta + n - y)$.

Find the posterior mean, density function and distribution function of $\psi = \theta^2$ in the case where $n = 5$, $y = 5$, and $\alpha = \beta = 1$.

Note: In the context where we toss a bent coin five times and get heads every time (and the prior on the probability of heads is standard uniform), the quantity ψ may be interpreted as the *probability of the next two tosses both coming up heads*, or equivalently, as the proportion of times heads will come up twice if the coin is repeatedly tossed in groups of two tosses a hypothetically infinite number of times.

Solution to Exercise 1.15

Here, $(\theta \mid y) \sim Beta(1+5, 1+5-5) \sim Beta(6,1)$
with pdf $f(\theta \mid y) = 6\theta^5, 0 < \theta < 1$.

Now $\theta = \psi^{1/2}$ and so, by the transformation method, the posterior density function of ψ is

$$f(\psi \mid y) = f(\theta \mid y)\left|\frac{d\theta}{d\psi}\right| = 6\psi^{5/2}\left|-\frac{1}{2}\psi^{-\frac{1}{2}}\right| = 3\psi^2, 0 < \psi < 1.$$

It follows that the posterior mean of ψ is

$$\hat{\psi} = E(\psi \mid y) = \int_0^1 \psi\left(3\psi^2\right)d\psi = 0.75,$$

and the posterior distribution function of ψ is

$$F(\psi \mid y) = \int_0^{\psi} f(\psi = t \mid y)dt = \int_0^{\psi} 3t^2 dt = \psi^3, 0 < \psi < 1.$$

Note 1: The posterior mean of $\psi = \theta^2$ can also be obtained by writing

$$\hat{\psi} = E(\theta^2 \mid y) = \int_0^1 \theta^2 \left(6\theta^5\right) d\theta = 0.75$$

or $$\hat{\psi} = E(\theta^2 \mid y) = V(\theta \mid y) + \{E(\theta \mid y)\}^2$$

$$= \frac{6 \times 1}{(6+1)^2(6+1+1)} + \left(\frac{6}{6+1}\right)^2 = 0.75$$

or $(\psi \mid y) \sim Beta(3,1) \quad \Rightarrow \quad \hat{\psi} = E(\psi \mid y) = 3/(3 + 1) = 0.75.$

Note 2: The distribution function of $\psi = \theta^2$ can also be obtained by writing

$$F(\psi = v \mid y) = P(\psi \le v \mid y) = P(\theta^2 \le v \mid y) = P(\theta \le v^{1/2} \mid y)$$

$$= F(\theta = v^{1/2} \mid y) = \left[\theta^6\Big|_{\theta = v^{1/2}}\right] = v^3,\ 0 < v < 1.$$

Note 3: In the above, $f(\psi = t \mid y)$ denotes the pdf of ψ given *y*, but evaluated at *t*. This pdf could also be written as $f_\psi(t \mid y)$ or as $\left[f(\psi \mid y)\Big|_{\psi = t}\right]$. Likewise, $F(\psi = v \mid y) \equiv F_\psi(v \mid y) \equiv \left[F(\psi \mid y)\Big|_{\psi = v}\right]$.

1.14 Credibility estimates

In actuarial studies, a *credibility estimate* is one which can be expressed as a weighted average of the form

$$C = (1-k)A + kB,$$

where:

- A is the *subjective estimate* (or the *collateral data estimate*)
- B is the *objective estimate* (or the *direct data estimate*)
- k is the *credibility factor*, a number that is between 0 and 1 (inclusive) and represents the weight assigned to the objective estimate.

A high value of *k* implies $C \cong B$, representing a situation where the objective estimate is assigned 'high credibility'. A primary aim of *credibility theory* is to determine an appropriate value or formula for *k*, as is done, for example, in the theory of the *Bühlmann model* (Bühlmann, 1967). Many Bayesian models lead to a point estimate which can be expressed as an intuitively appealing credibility estimate.

Exercise 1.16 Credibility estimation in the binomial-beta model

Consider the binomial-beta model: $(y\,|\,\theta) \sim Binomial(n,\theta)$
$\theta \sim Beta(\alpha,\beta)$.

Express the posterior mean of θ as a credibility estimate and discuss.

Solution to Exercise 1.16

Earlier we showed that

$$(\theta\,|\,y) \sim Beta(\alpha+y,\beta+n-y),$$

and hence that the posterior mean of θ is

$$\hat{\theta} = E(\theta\,|\,y) = \frac{(\alpha+y)}{(\alpha+y)+(\beta+n-y)} = \frac{\alpha+y}{\alpha+\beta+n}.$$

Observe that the prior mean of θ is $E\theta = \alpha/(\alpha+\beta)$, and the maximum likelihood estimate (MLE) of θ is y/n. This suggests that we write

$$\begin{aligned}\hat{\theta} &= \frac{\alpha}{\alpha+\beta+n} + \frac{y}{\alpha+\beta+n}\\ &= \frac{\alpha}{\alpha+\beta+n}\left(\frac{\alpha+\beta}{\alpha}\right)\left(\frac{\alpha}{\alpha+\beta}\right) + \frac{n}{\alpha+\beta+n}\left(\frac{y}{n}\right)\\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right) + \frac{n}{\alpha+\beta+n}\left(\frac{y}{n}\right).\end{aligned}$$

Thus $\hat{\theta} = (1-k)A + kB$

where: $A = \dfrac{\alpha}{\alpha+\beta}$, $B = \dfrac{y}{n}$, $k = \dfrac{n}{\alpha+\beta+n}$.

We see that the posterior mean $\hat{\theta}$ is a credibility estimate in the form of a weighted average of the prior mean $A = E\theta = \alpha/(\alpha+\beta)$ and the MLE $B = y/n$, where the weight assigned to the MLE is the credibility factor given by $k = n/(n+\alpha+\beta)$. Observe that as n increases, the credibility factor k approaches 1. This makes sense: if there is a lot of data then the prior should not have much influence on the estimation.

Figure 1.9 illustrates this idea by showing relevant densities, likelihoods and estimates for the following two cases, respectively:

(a) $n = 5, \ \ y = 4, \ \ \alpha = 2, \ \beta = 6$

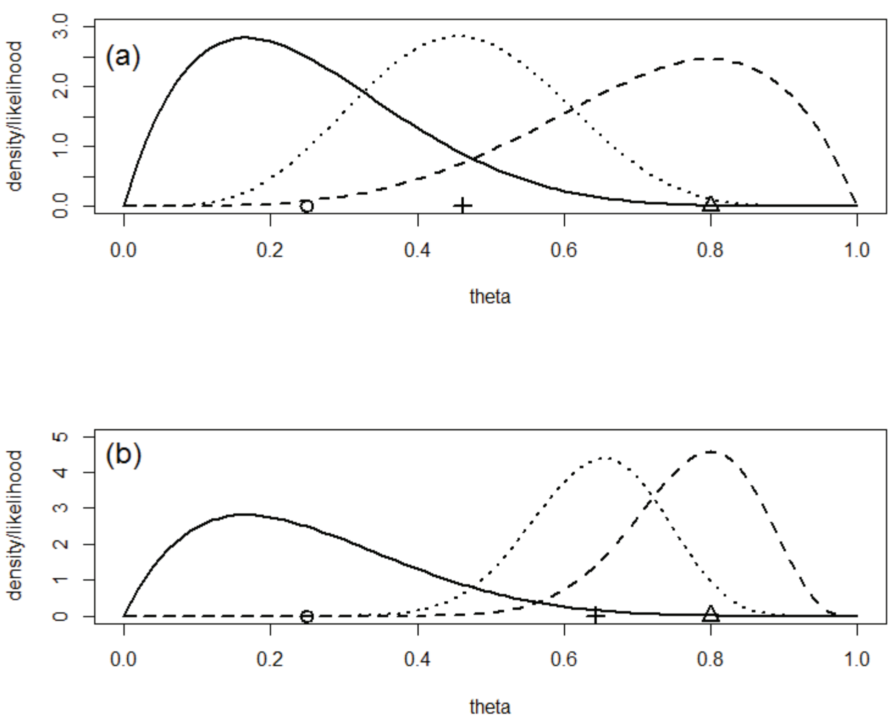
(b) $n = 20, y = 16, \ \alpha = 2, \ \beta = 6.$

In both cases, the prior mean is the same (A = 2/(2 + 6) = 0.25), as is the MLE (B = 4/5 = 16/20 = 0.8). However, due to n being larger in case (b) (i.e. there being more direct data), case (b) leads to a larger credibility factor (0.714 compared to 0.385) and hence a posterior mean closer to the MLE (0.643 compared to 0.462).

Note: Each likelihood function in Figure 1.9 has been normalised so that the area underneath it is exactly 1. This means that in each case (a) and (b), the likelihood function $L(\theta)$ as shown is identical to the posterior density which would be implied by the standard uniform prior, i.e. under $f_{U(0,1)}(\theta) = f_{Beta(1,1)}(\theta)$. Thus, $L(\theta) = f_{Beta(1+y,1+n-y)}(\theta)$.

Figure 1.9 Illustration for Exercise 1.16

Legend: solid line = prior, dashed line = likelihood, dotted line = posterior, circle = prior mean, triangle = MLE, cross = posterior mean



R Code for Exercise 1.16

```
X11(w=8,h=7); par(mfrow=c(2,1))

alp=2; bet=6; n = 5; y = 4; pvec=seq(0,1,0.01)
plot(c(0,1),c(0,3),type="n",xlab="theta",ylab="density/likelihood")
lines(pvec,dbeta(pvec,alp,bet),lty=1,lwd=2)
lines(pvec,dbeta(pvec,1+y,n-y+1),lty=2,lwd=2)
lines(pvec,dbeta(pvec,alp+y,n-y+bet),lty=3,lwd=2)

points(c(alp/(alp+bet), y/n,(alp+y)/(alp+bet+n)),c(0,0,0),pch=c(1,2,3),
        cex=rep(1.5,3),lwd=2);  text(0,2.5,"(a)",cex=1.5)
c(alp/(alp+bet), y/n,(alp+y)/(alp+bet+n)) # 0.2500000 0.8000000 0.4615385
n/(alp+bet+n) # 0.3846154

alp=2; bet=6; n = 20; y = 16; pvec=seq(0,1,0.01)
plot(c(0,1),c(0,5),type="n",xlab="theta",ylab="density/likelihood")
lines(pvec,dbeta(pvec,alp,bet),lty=1,lwd=2)
lines(pvec,dbeta(pvec,1+y,n-y+1),lty=2,lwd=2)
lines(pvec,dbeta(pvec,alp+y,n-y+bet),lty=3,lwd=2)

points(c(alp/(alp+bet), y/n,(alp+y)/(alp+bet+n)),c(0,0,0),pch=c(1,2,3),
        cex=rep(1.5,3),lwd=2);  text(0,4.5,"(b)",cex=1.5)
c(alp/(alp+bet), y/n,(alp+y)/(alp+bet+n)) # 0.2500000 0.8000000 0.6428571
n/(alp+bet+n) # 0.7142857
```

Exercise 1.17 Further credibility estimation in the binomial-beta model

Consider the binomial-beta model:

$$(Y \mid \theta) \sim Binomial(n, \theta)$$
$$\theta \sim Beta(\alpha, \beta).$$

If possible, express the posterior *mode* of θ as a credibility estimate.

Solution to Exercise 1.17

Since $(\theta \mid y) \sim Beta(\alpha + y, \beta + n - y)$, the posterior mode of θ is

$$Mode(\theta \mid y) = \frac{(\alpha + y - 1)}{(\alpha + y - 1) + (\beta + n - y - 1)} = \frac{\alpha + y - 1}{\alpha + \beta + n - 2}.$$

Now, the prior mode of θ is $Mode(\theta)=\dfrac{(\alpha-1)}{(\alpha-1)+(\beta-1)}=\dfrac{\alpha-1}{\alpha+\beta-2}$.

So we write $Mode(\theta\,|\,y)=\dfrac{\alpha-1}{\alpha+\beta+n-2}+\dfrac{y}{\alpha+\beta+n-2}$

$$=\frac{\alpha-1}{\alpha+\beta+n-2}\left(\frac{\alpha+\beta-2}{\alpha-1}\right)\left(\frac{\alpha-1}{\alpha+\beta-2}\right)+\frac{n}{\alpha+\beta+n-2}\left(\frac{y}{n}\right).$$

We see that the posterior mode is a credibility estimate of the form

$$Mode(\theta\,|\,y)=(1-c)Mode(\theta)+c\hat{\theta},$$

where: $Mode(\theta)=\dfrac{\alpha-1}{\alpha+\beta-2}$ is the prior mode

$\hat{\theta}=\dfrac{y}{n}$ is the maximum likelihood estimate

(mode of the likelihood function)

$c=\dfrac{n}{n+\alpha+\beta-2}$ is the credibility factor

(assigned to the direct data estimate, $\hat{\theta}$).

Exercise 1.18 The *normal-normal model*

Consider the following Bayesian model:

$$(y_1,\dots,y_n\,|\,\mu)\sim iid\ N(\mu,\sigma^2)$$

$$\mu\sim N(\mu_0,\sigma_0^2),$$

where σ^2, μ_0 and σ_0^2 are known or specified constants.

Find the posterior distribution of μ given data in the form of the vector $y=(y_1,...,y_n)$.

Solution to Exercise 1.18

The posterior density of μ is

$$f(\mu\,|\,y)\propto f(\mu)f(y\,|\,\mu)$$

$$\overset{\mu}{\propto}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\times\prod_{i=1}^{n}\exp\left\{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right\}$$

$$= \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}\left(\mu^2 - 2\mu\mu_0 + \mu_0^2\right) + \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\mu n\bar{y} + n\mu^2\right)\right]\right), \quad (1.1)$$

where $\bar{y} = (y_1 + ... + y_n)/n$ is the sample mean.

We see that the posterior density of μ is proportional to the exponent of a quadratic in μ. That is,

$$f(\mu \mid y) \propto \exp\left(-\frac{1}{2\sigma_*^2}(\mu - \mu_*)^2\right), \quad (1.2)$$

which then implies that

$$(\mu \mid y) \sim N(\mu_*, \sigma_*^2),$$

for some constants μ_* and σ_*^2.

It remains to find the normal mean and variance parameters, μ_* and σ_*^2. (These must be functions of the known quantities n, $\bar{y}$, σ, μ_0 and σ_0.)

One way to obtain these parameters which completely define μ's posterior distribution is to complete the square in the exponent of (1.2). To this end we write

$$f(\mu \mid y) \propto \exp\left(-\frac{1}{2}q\right),$$

where

$$q = \frac{1}{\sigma_0^2}\left(\mu^2 - 2\mu\mu_0\right) + \frac{1}{\sigma^2}\left(-2\mu n\bar{y} + n\mu^2\right)$$

(ignoring constants with respect to μ)

$$= \mu^2\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right) + c$$

(where c is a constant with respect to μ)

$$= a\mu^2 - 2b\mu + c \quad \text{where} \quad a = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \quad \text{and} \quad b = \frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}$$

$$= a\left(\mu^2 - 2\frac{b}{a}\mu\right) + c \quad = a\left(\mu^2 - 2\left(\frac{b}{a}\right)\mu + \left(\frac{b}{a}\right)^2\right) + c'$$

(where c' is a constant with respect to μ)

$$= \frac{1}{1/a}\left(\mu - \frac{b}{a}\right)^2 + c'.$$

Thus, $$f(\mu \mid y) \propto \exp\left(-\frac{1}{2(1/a)}\left(\mu - \frac{b}{a}\right)^2\right). \tag{1.3}$$

So, equating (1.2) and (1.3), we obtain:

$$\sigma_*^2 = \frac{1}{a} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_* = \frac{b}{a} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2\mu_0 + n\sigma_0^2\bar{y}}{\sigma^2 + n\sigma_0^2}. \tag{1.4}$$

Note 1: A little algebra (left as an additional exercise) shows that the posterior mean can also be written as

$$\mu_* = (1-k)\mu_0 + k\bar{y},$$

and the posterior variance can be written as

$$\sigma_*^2 = k\frac{\sigma^2}{n},$$

where

$$k = \frac{n}{n + \frac{\sigma^2}{\sigma_0^2}}.$$

We see that μ's posterior mean is a credibility estimate in the form of a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$ (which is also the maximum likelihood estimate), with the weight assigned to $\bar{y}$ being the credibility factor, k. More will be said on this further down.

Note 2: Another way to derive μ_* and σ_*^2 is to write (1.2) as

$$f(\mu \mid y) \propto \exp\left(-\frac{1}{2\sigma_*^2}\left(\mu^2 - 2\mu\mu_* + \mu_*^2\right)\right) \tag{1.5}$$

and then equate coefficients of powers of μ in (1.1) and (1.5). This logic leads to $\frac{1}{\sigma_*^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$ and $\frac{\mu_*}{\sigma_*^2} = \frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}$ and ultimately the same formulae for μ_* and σ_*^2 as given by (1.4).

Note 3: Since both prior and posterior are normal, the prior is *conjugate*.

Note 4: The posterior mean, mode and median of μ are the same and equal to μ_*. The $1-\alpha$ CPDR and $1-\alpha$ HPDR for μ are the same and equal to $(\mu_* \pm z_{\alpha/2}\sigma_*)$.

Note 5: The posterior distribution of μ depends on the data vector $y=(y_1,\ldots,y_n)$ only by way of the sample mean, i.e. $\bar{y}=(y_1+\ldots+y_n)/n$. Therefore, the main result, $(\mu \mid y) \sim N(\mu_*,\sigma_*^2)$, also implies that $(\mu \mid \bar{y}) \sim N(\mu_*,\sigma_*^2)$.

That is, if we know only the sample mean $\bar{y}$, the posterior distribution of μ is the same as if we know *y*, i.e. all *n* sample values. Knowing the individual y_i values makes no difference to the inference.

Note 6: The formula for the credibility factor in Note 1, namely

$$k=\frac{n}{n+\frac{\sigma^2}{\sigma_0^2}}=\frac{1}{1+\frac{\sigma^2/n}{\sigma_0^2}},$$

makes sense in the following ways:

(i) If the prior standard deviation σ_0 is *small* then $k \approx 0$, so that $\mu_* \approx \mu_0$ and $\sigma_* \approx \sigma_0$. Therefore $(\mu \mid y) \mathrel{\dot\sim} N(\mu_0,\sigma_0^2)$.

That is, if the prior information is very ‘precise’ or ‘definite’, the data has little influence on the posterior. So the posterior is approximately equal to the prior; i.e. $f(\mu \mid y) \approx f(\mu)$, or equivalently, $(\mu \mid y) \mathrel{\dot\sim} \mu$. In this case the posterior mean, mode and median of μ are approximately equal to μ_0. Also, the $1-\alpha$ CPDR and $1-\alpha$ HPDR for μ are approximately equal to $(\mu_0 \pm z_{\alpha/2}\sigma_0)$.

(ii) If σ_0 is *large* then $k \approx 1$, so that $\mu_* \approx \bar{y}$, $\sigma_*^2 \approx \sigma^2/n$, and so $(\mu \mid y) \mathrel{\dot\sim} N(\bar{y},\sigma^2/n)$.

That is, a large σ_0 corresponds to a highly disperse prior, reflecting little prior information and so little influence of the prior distribution (as specified by μ_0 and σ_0) on the inference. In this case the posterior mean, mode and median of μ are approximately equal to $\bar{y}$. Also, the $1-\alpha$ CPDR and $1-\alpha$ HPDR for μ are approximately equal to $(\bar{y} \pm z_{\alpha/2}\sigma / \sqrt{n})$. Thus, inference is almost the same as implied by the classical approach.

(iii) If the sample size n is *large* then $k \approx 1$, so that $\mu_* \approx \bar{y}$ and $\sigma_*^2 \approx \sigma^2 / n$. Therefore $(\mu \mid y) \mathrel{\dot\sim} N(\bar{y}, \sigma^2 / n)$.

So, in this case, just as when σ_0 is large, the prior distribution has very little influence on the posterior, and the ensuing inference is almost the same as that implied by the classical approach.

Note 7: In the case of *a priori* ignorance (meaning no prior information at all) it is customary to take $\sigma_0 = \infty$, which implies that

$$\mu \sim N(0, \infty).$$

This prior on μ appears to be problematic, because it is *improper*. However, it meaningfully leads to a *proper* posterior, namely

$$(\mu \mid y) \sim N(\bar{y}, \sigma^2 / n),$$

which then leads to the same point and interval estimates implied by the classical approach, namely the MLE $\bar{y}$ and $1-\alpha$ CI $(\bar{y} \pm z_{\alpha/2}\sigma / \sqrt{n})$.

The improper prior $\mu \sim N(0, \infty)$ may be described as ‘flat’ or ‘uniform over the whole real line’ and can also be written as

$$\mu \sim U(-\infty, \infty)$$

or $\quad f(\mu) \propto 1, \ \mu \in \Re$.

In some cases (more complicated models not considered here), using an improper prior may lead to an improper posterior, which then becomes problematic. For more information on this topic, see Hobert and Casella (1996).

Summary: For the *normal-normal model*, defined by:

$$(y_1, \ldots, y_n \mid \mu) \sim iid\ N(\mu, \sigma^2)$$
$$\mu \sim N(\mu_0, \sigma_0^2),$$

the posterior distribution of the normal mean μ is given by

$$(\mu \mid y) \sim N(\mu_*, \sigma_*^2),$$

where: $\mu_* = (1-k)\mu_0 + k\bar{y}$

$$\sigma_*^2 = k\frac{\sigma^2}{n}$$

$$k = \frac{n}{n + \sigma^2 / \sigma_0^2}$$ (the *normal-normal model credibility factor*).

The posterior mean, mode and median of μ are all equal to μ_*,
and the $1-\alpha$ CPDR and HPDR for μ are both $(\mu_* \pm z_{\alpha/2}\sigma_*)$.

In the case of *a priori* ignorance it is appropriate to set $\sigma_0 = \infty$.

This defines an improper prior

$$f(\mu) \propto 1, \mu \in \Re$$

and the proper posterior

$$(\mu \mid y) \sim N(\bar{y}, \sigma^2 / n).$$

Exercise 1.19 Practice with the normal-normal model

In the context of the normal-normal model, given by:

$$(y_1, \ldots, y_n \mid \mu) \sim iid\ N(\mu, \sigma^2)$$
$$\mu \sim N(\mu_0, \sigma_0^2),$$

suppose that y = (8.4, 10.1, 9.4), σ = 1, μ_0 = 5 and σ_0 = 1/2.

Calculate the posterior mean, mode and median of μ.

Also calculate the 95% CPDR and 95% HPDR for μ.

Create a graph which shows these estimates as well as the prior density, prior mean, likelihood, MLE and posterior density.

Solution to Exercise 1.19

Here: $n = 3$,

$\bar{y} = (8.4 + 10.1 + 9.4)/3 = 9.3$

$$k = \frac{1}{1 + \frac{1^2/3}{(1/2)^2}} = \frac{3}{7} = 0.4285714$$

$$\mu_* = \left(1 - \frac{3}{7}\right)5 + \frac{3}{7} \times 9.3 = 6.8428571$$

$$\sigma_*^2 = \frac{3}{7} \times \frac{1^2}{3} = \frac{1}{7} = 0.1428571.$$

So the posterior mean/mode/median is

$\mu_* = 6.84286$,

and the 95% CPDR/HPDR is

$$(\mu_* \pm z_{0.025}\sigma_*) = (6.84286 \pm 1.96\sqrt{0.14286})$$
$$= (6.102,\ 7.584).$$

Figure 1.10 shows the various densities and estimates here, as well as the normalised likelihood. Note that the likelihood function as shown is also the posterior density if the prior is taken to be uniform over the whole real line, i.e. $\mu \sim U(-\infty, \infty)$.

Discussion

If we change σ_0 from 0.5 to 2 we get k = 0.923 and results as illustrated in Figure 1.11.

If we change σ_0 from 0.5 to 0.25 we get k = 0.158 and results as illustrated in Figure 1.12 (page 46).

If we keep σ_0 as 0.5 but change σ from 1 to 2 we get k = 0.158 and results as illustrated in Figure 1.13 (page 46).

Note that the posteriors in Figures 1.12 and 1.13 have the same mean but different variances.

Figure 1.10 Results if $\sigma_0 = 0.5,\ \sigma = 1,\ k = n/(n + \sigma^2/\sigma_0^2) = 0.429$

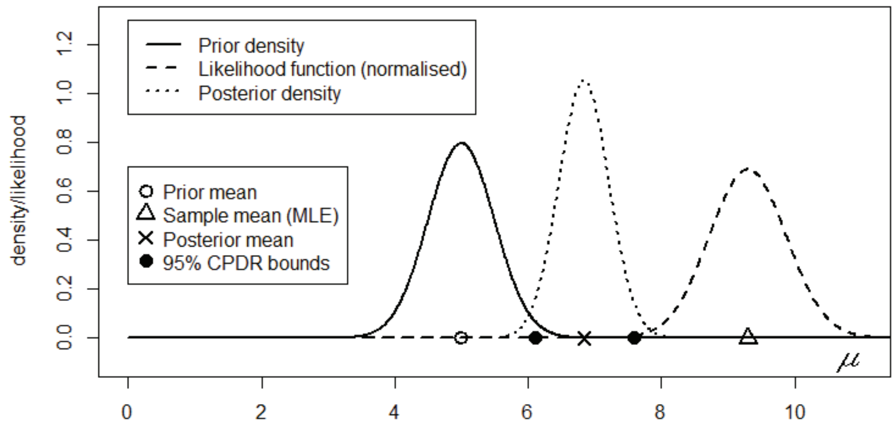


Figure 1.11 Results if $\sigma_0 = 2,\ \sigma = 1,\ k = n/(n + \sigma^2/\sigma_0^2) = 0.9223$

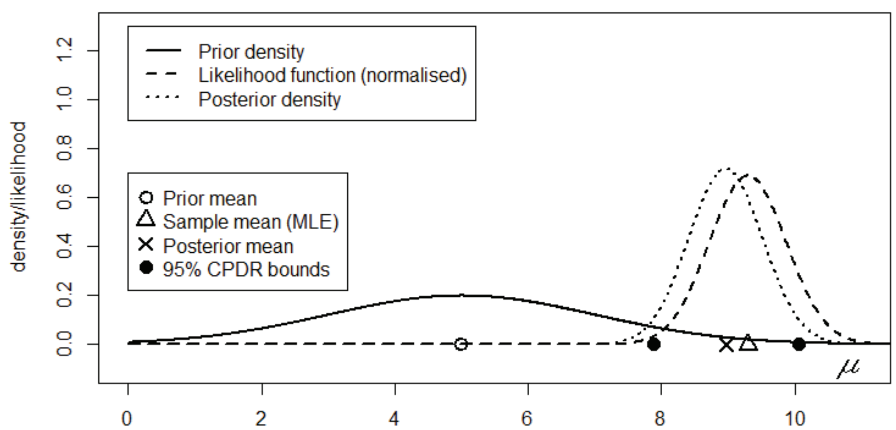


Figure 1.12 Results if $\sigma_0 = 0.25$, $\sigma = 1$, $k = n/(n + \sigma^2/\sigma_0^2) = 0.158$

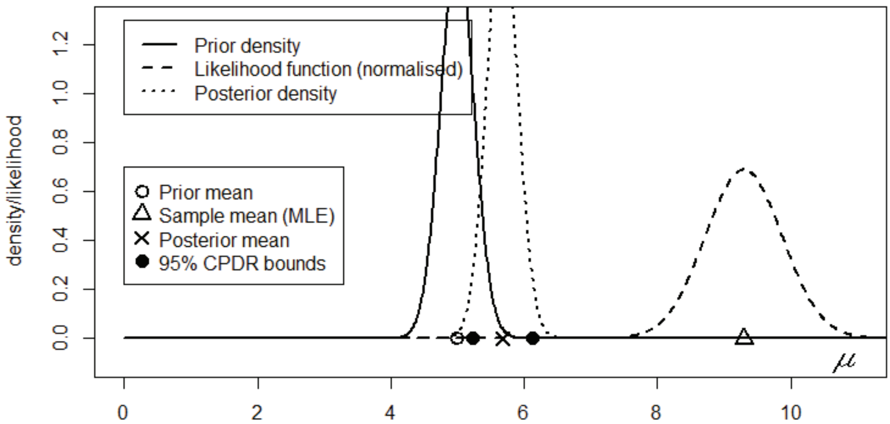
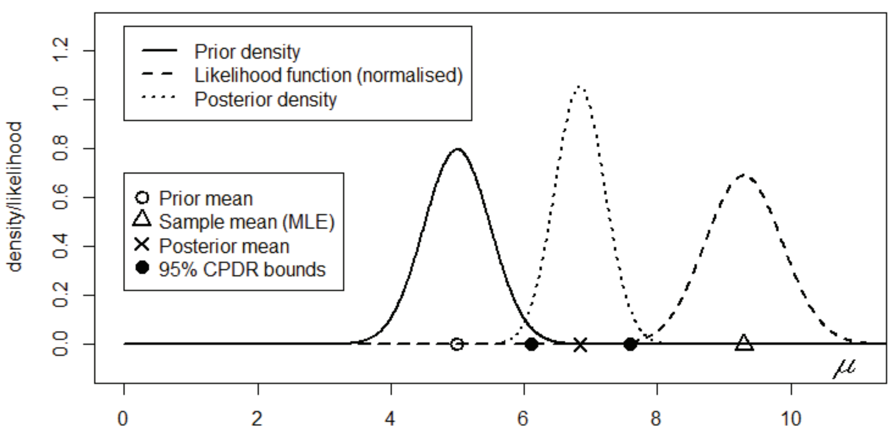


Figure 1.13 Results if $\sigma_0 = 0.5$, $\sigma = 2$, $k = n/(n + \sigma^2/\sigma_0^2) = 0.158$



R Code for Exercise 1.19

```
X11(w=8,h=5); par(mfrow=c(1,1)); mu0=5;  sig0=0.5; sig=1

y = c(8.4, 10.1, 9.4); n = length(y); k=1/(1+(sig^2/n)/sig0^2); k # 0.4285714
ybar=mean(y); ybar # 9.3
mus = (1-k)*mu0 + k*ybar; sigs2=k*sig^2/n
c(mus,sigs2) # 6.8428571 0.1428571
muv=seq(0,15,0.01)
prior = dnorm(muv,mu0,sig0); post=dnorm(muv,mus,sqrt(sigs2))
like = dnorm(muv,ybar,sig/sqrt(n))
cpdr=mus+c(-1,1)*qnorm(0.975)*sqrt(sigs2)
cpdr # 6.102060 7.583654

plot(c(0,11),c(-0.1,1.3),type="n",xlab="",ylab="density/likelihood")
lines(muv,prior,lty=1,lwd=2); lines(muv,like,lty=2,lwd=2)
lines(muv,post,lty=3,lwd=2)
points(c(mu0,ybar,mus),c(0,0,0),pch=c(1,2,4),cex=rep(1.5,3),lwd=2)
points(cpdr,c(0,0),pch=rep(16,2),cex=rep(1.5,2))
legend(0,1.3,
    c("Prior density","Likelihood function (normalised)","Posterior density"),
        lty=c(1,2,3),lwd=c(2,2,2))
legend(0,0.7,c("Prior mean","Sample mean (MLE)","Posterior mean",
   "95% CPDR bounds"), pch=c(1,2,4,16),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))
text(10.8,-0.075,"m", vfont=c("serif symbol","italic"), cex=1.5)

# Repeat above with sig0=2 to obtain Figure 1.11
# Repeat above with sig0=0.25 to obtain Figure 1.12
# Repeat above with sig0=0.5 and sig=2 to obtain Figure 1.13
```

Exercise 1.20 The *normal-gamma model*

Consider the following Bayesian model:

$$(y_1,\ldots,y_n \mid \lambda) \sim iid\ N(\mu, 1/\lambda)$$
$$\lambda \sim G(\alpha,\beta)\,.$$

Find the posterior distribution of λ given $y=(y_1,...,y_n)$.

Note 1: In the normal-normal model, the normal mean μ is unknown and the normal variance σ^2 is known. Now we consider the same Bayesian model but with those roles reversed, i.e. with μ known and σ^2 unknown. For an example of where this kind of situation might arise, see Byrne and Dracoulis (1985).

Note 2: For reasons of mathematical convenience and conjugacy, we parameterise the normal distribution here via the *precision parameter*

$$\lambda = 1/\sigma^2$$

rather than using σ^2 directly as before in the normal-normal model.

Note 3: An equivalent formulation of the *normal-gamma model* being considered here is:

$$(y_1,\ldots,y_n \mid \sigma^2) \sim iid\ N(\mu,\sigma^2)$$

$$\sigma^2 \sim IG(\alpha,\beta),$$

where this may be called the *normal-inverse-gamma* model.

Solution to Exercise 1.20

The posterior density of λ is

$$\begin{aligned} f(\lambda \mid y) &\propto f(\lambda) f(y \mid \lambda) \\ &\overset{\lambda}{\propto} \lambda^{\alpha-1}e^{-\beta\lambda} \times \prod_{i=1}^{n} \frac{1}{1/\sqrt{\lambda}} \exp\left\{-\frac{1}{2}\left(\frac{y_i-\mu}{1/\sqrt{\lambda}}\right)^2\right\} \\ &= \lambda^{\alpha-1}e^{-\beta\lambda} \times \lambda^{n/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \\ &= \lambda^{a-1}e^{-b\lambda} \quad \text{for some } a \text{ and } b. \end{aligned}$$

We see that

$$(\lambda \mid y) \sim G(a,b),$$

where: $a = \alpha + \frac{n}{2}$

$$b = \beta + \frac{n}{2}s_\mu^2$$

$$s_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 .$$

Note 1: The posterior mean of λ, namely

$$E(\lambda \mid y) = \frac{a}{b} = \frac{\alpha + n/2}{\beta + ns_\mu^2/2},$$

converges to $\hat{\lambda} = \frac{1}{s_\mu^2}$ (the MLE of λ) as $n \to \infty$.

If $\alpha = \beta = 0$ then $E(\lambda \mid y) = \hat{\lambda}$ exactly for all *n*.

Note 2: Unlike the posterior mean of μ in the normal-normal model, the posterior mean of λ cannot be expressed as a credibility estimate of the form

$$(1-c)\lambda_0 + c\hat{\lambda},$$

where: $\lambda_0 = E\lambda = \frac{\alpha}{\beta}$ (the prior mean of λ)

$\hat{\lambda} = \frac{1}{s_\mu^2}$ (the MLE of λ).

Note 3: We may write the posterior as

$$(\lambda \mid y) \sim G\left(\frac{2\alpha + n}{2}, \frac{2\beta + ns_\mu^2}{2}\right).$$

It can then be shown via the method of transformations that

$$(u \mid y) \sim G\left(\frac{2\alpha + n}{2}, \frac{1}{2}\right) \sim \chi^2(2\alpha + n),$$

where $u = (2\beta + ns_\mu^2)\lambda$.

So the $1-A$ CPDR for *u* is $\left(\chi^2_{1-A/2}(2\alpha+n), \chi^2_{A/2}(2\alpha+n)\right)$.

So the $1-A$ CPDR for $\lambda = \frac{u}{2\beta + ns_\mu^2}$ is $\left(\frac{\chi^2_{1-A/2}(2\alpha+n)}{2\beta + ns_\mu^2}, \frac{\chi^2_{A/2}(2\alpha+n)}{2\beta + ns_\mu^2}\right)$.

So the $1-A$ CPDR for $\sigma^2 = \frac{1}{\lambda}$ is $\left(\frac{2\beta + ns_\mu^2}{\chi^2_{A/2}(2\alpha+n)}, \frac{2\beta + ns_\mu^2}{\chi^2_{1-A/2}(2\alpha+n)}\right)$.

If $\alpha = \beta = 0$, this is exactly the same as the classical $1-A$ CI for σ^2.

Note 4: The classical $1-A$ CI for σ^2 may be derived as follows. First consider all parameters fixed as constants. Then

$$\frac{y_1-\mu}{\sigma},\ldots,\frac{y_n-\mu}{\sigma} \sim iid\ N(0,1).$$

So

$$\left(\frac{y_1-\mu}{\sigma}\right)^2,\ldots,\left(\frac{y_n-\mu}{\sigma}\right)^2 \sim iid\ \chi^2(1).$$

So

$$\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2 = \frac{ns_\mu^2}{\sigma^2} \sim \chi^2(n).$$

So

$$1-A = P\left(\chi^2_{1-A/2}(n) < \frac{ns_\mu^2}{\sigma^2} < \chi^2_{A/2}(n)\right)$$
$$= P\left(\frac{ns_\mu^2}{\chi^2_{A/2}(n)} < \sigma^2 < \frac{ns_\mu^2}{\chi^2_{1-A/2}(n)}\right).$$

Note 5: Notes 1 to 3 indicate that in the case of *a priori* ignorance, a reasonable specification is

$$\alpha = \beta = 0,$$

or equivalently,

$$f(\lambda) \propto 1/\lambda, \quad \lambda > 0.$$

This improper prior may be thought of as the limiting case as $\varepsilon \to 0$ of the proper prior

$$\lambda \sim \text{Gam}(\varepsilon,\varepsilon),$$

where $\varepsilon \approx 0$.

Observe that

$$E\lambda = \varepsilon/\varepsilon = 1$$

for all ε, and

$$V\lambda = \varepsilon/\varepsilon^2 \to \infty$$

as $\varepsilon \to 0$.

Summary: For the *normal-gamma model*, defined by:

$$(y_1,\ldots,y_n \mid \lambda) \sim iid\ N(\mu, 1/\lambda)$$
$$\lambda \sim G(\alpha,\beta),$$

the posterior distribution of λ is given by

$$(\lambda \mid y) \sim G(a,b),$$

where: $a = \alpha + \dfrac{n}{2}$, $b = \beta + \dfrac{n}{2} s_\mu^2$, $s_\mu^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$.

The posterior mean of λ is *a*/*b*. The posterior median is $F^{-1}_{G(a,b)}(1/2)$.
The posterior mode of λ is $(a-1)/b$ if $a > 1$; otherwise that mode is 0.

The $1-A$ CPDR for λ is $\left(F^{-1}_{G(a,b)}(A/2), F^{-1}_{G(a,b)}(1-A/2)\right)$
and may also be written as $\left(\dfrac{\chi^2_{1-A/2}(2\alpha+n)}{2\beta+ns_\mu^2}, \dfrac{\chi^2_{A/2}(2\alpha+n)}{2\beta+ns_\mu^2}\right)$.

The $1-A$ CPDR for $\sigma^2 = 1/\lambda$ is $\left(\dfrac{2\beta+ns_\mu^2}{\chi^2_{A/2}(2\alpha+n)}, \dfrac{2\beta+ns_\mu^2}{\chi^2_{1-A/2}(2\alpha+n)}\right)$.

In the case of *a priori* ignorance it is appropriate to set $\alpha = \beta = 0$.
This defines an improper prior with density

$$f(\lambda) \propto 1/\lambda, \lambda > 0,$$

and a proper posterior distribution given by

$$(ns_\mu^2 \lambda \mid y) \sim \chi^2(n).$$

Exercise 1.21 Practice with the normal-gamma model

In the context of the normal-gamma model, given by:

$$(y_1,\ldots,y_n \mid \lambda) \sim iid\ N(\mu, 1/\lambda)$$
$$\lambda \sim Gamma(\alpha,\beta),$$

suppose that $y = (8.4, 10.1, 9.4)$, $\mu = 8$, $\alpha = 3$ and $\beta = 2$.

(a) Calculate the posterior mean, mode and median of the model precision λ. Also calculate the 95% CPDR for λ. Create a graph which shows these estimates as well as the prior density, prior mean, likelihood, MLE and posterior density.

(b) Calculate the posterior mean, mode and median of the model variance $\sigma^2 = 1/\lambda$. Also calculate the 95% CPDR for σ^2. Create a graph which shows these estimates as well as the prior density, prior mean, likelihood, MLE and posterior density.

(c) Calculate the posterior mean, mode and median of the model standard deviation σ. Also calculate the 95% CPDR for σ. Create a graph which shows these estimates as well as the prior density, prior mean, likelihood, MLE and posterior density.

(d) Examine each of the point estimates in (a), (b) and (c) and determine which ones, if any, can be easily expressed in the form of a credibility estimate.

Solution to Exercise 1.21

(a) The required posterior distribution is $(\lambda \mid y) \sim Gamma(a,b)$, where:

$$a = \alpha + \frac{n}{2} = 4.5, \; b = \beta + \frac{n}{2} s_\mu^2 = 5.265, \; s_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2 = 2.177.$$

So:

- the posterior mean of λ is $E(\lambda \mid y) = a/b = 0.8547$
- the posterior mode is $Mode(\lambda \mid y) = (a-1)/b = 0.6648$
- the posterior median is the 0.5 quantile of the $G(a,b)$ distribution and works out as $Median(\lambda \mid y) = 0.7923$ (as obtained using the qgamma() function in R; see below)
- the 95% CPDR for λ is (0.2564, 1.8065) (where the bounds are the 0.025 and 0.975 quantiles of the $G(a,b)$ distribution).

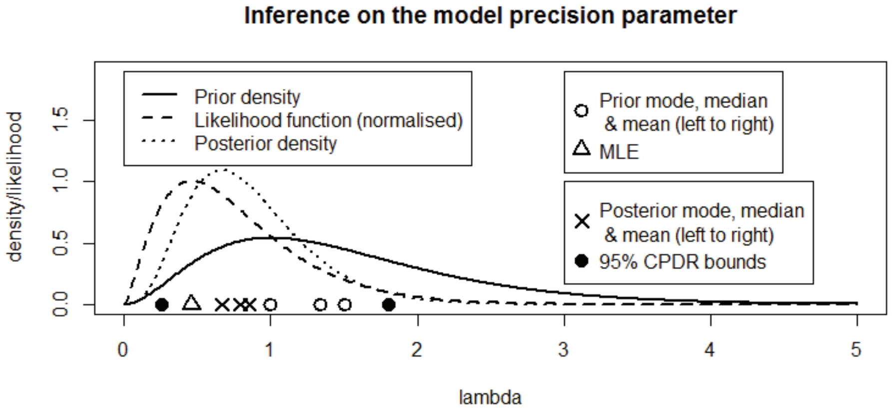
Also:

- the prior mean is $E\lambda = \alpha/\beta = 1.5$
- the prior mode is $Mode(\lambda) = (\alpha - 1)/\beta = 1$
- the prior median is $Median(\lambda) = 1.3370$
- the MLE of λ is $\hat{\lambda} = 1/s_\mu^2 = 0.4594$ (note that this estimate is biased).

Figure 1.14 shows the various densities and estimates here, as well as the normalised likelihood function.

Note: The normalised likelihood function (with area below equal to 1) is the same as the posterior density of λ if the prior is taken to be uniform over the positive real line, i.e. $\lambda \sim U(0,\infty)$. This prior is specified by taking $\alpha = 1$ and $\beta = 0$, because then $f(\lambda) \propto \lambda^{1-1} e^{-0\lambda} \propto 1$.

Figure 1.14 Results for Exercise 1.21(a)



(b) As regards the model variance $\sigma^2 = 1/\lambda$ we note that $\sigma^2 \sim IG(\alpha,\beta)$ with density

$$f(\sigma^2) = f(\lambda)\left|\frac{d\lambda}{d\sigma^2}\right| \quad \text{where } \lambda = \left(\sigma^2\right)^{-1}$$

$$= \frac{\beta^\alpha [(\sigma^2)^{-1}]^{\alpha-1} e^{-\beta(\sigma^2)^{-1}}}{\Gamma(\alpha)} \left|-(\sigma^2)^{-2}\right|$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2}, \sigma^2 > 0. \qquad (1.6)$$

Then, by well-known properties of the inverse gamma distribution and maximum likelihood theory:

- the prior mean of σ^2 is $E\sigma^2 = \beta/(\alpha-1) = 1$
- the prior mode is $Mode(\sigma^2) = \beta/(\alpha+1) = 0.5$
- the prior median is $Median(\sigma^2) = 1/Median(\lambda) = 0.7479$
- the MLE of σ^2 is $\hat{\sigma}^2 = 1/\hat{\lambda} = s_\mu^2 = 2.1767$
  (note that this estimate is unbiased).

By analogy with the prior (1.6), we find that $(\sigma^2 \mid y) \sim IG(a,b)$ with density

$$f(\sigma^2 \mid y) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} e^{-b/\sigma^2}, \sigma^2 > 0,$$

and hence that:

- the posterior mean of σ^2 is $E(\sigma^2 \mid y) = b/(a-1)$ = 1.5043
- the posterior mode is $Mode(\sigma^2 \mid y) = b/(a+1)$ = 0.9573
- the posterior median is

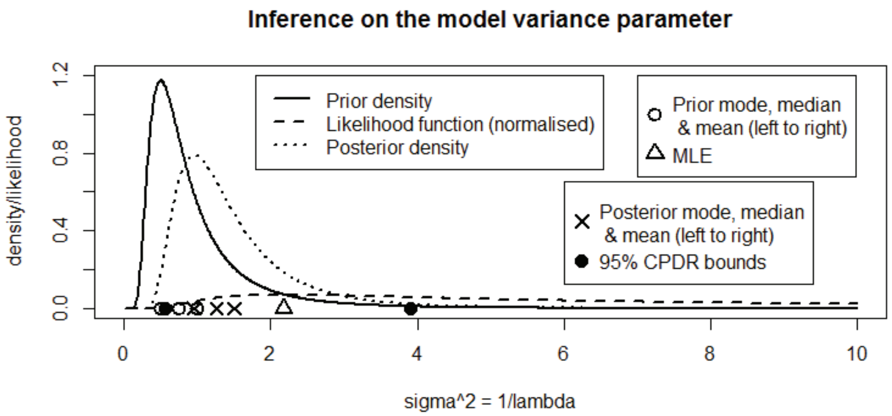
  $$Median(\sigma^2 \mid y) = 1/Median(\lambda \mid y) = 1.2622$$

  (since $1/2 = P(\sigma^2 < m \mid y) = P(1/\lambda < m \mid y) = P(1/m < \lambda \mid y)$)
- the 95% CPDR for σ^2 is (0.5535, 3.8994) (where the lower and upper bounds are the inverses of the 0.975 and 0.025 quantiles of the $G(a,b)$ distribution, respectively).

Figure 1.15 shows the various densities and estimates here, as well as the normalised likelihood function.

Note: The normalised likelihood function is the same as the posterior density of σ^2 if the prior on σ^2 is taken to be uniform over the positive real line, i.e. $f(\sigma^2) \propto 1, \sigma^2 > 0$. This prior is specified by $\lambda \sim G(-1,0)$, i.e. by $\alpha = -1$ and $\beta = 0$ as is evident from (1.6) above.

Figure 1.15 Results for Exercise 1.21(b)



(c) As regards the model standard deviation $\sigma = 1/\sqrt{\lambda}$, observe that the prior density of this quantity is

$$f(\sigma) = f(\lambda)\left|\frac{d\lambda}{d\sigma}\right| \quad \text{where } \lambda = \sigma^{-2}$$

$$= \frac{\beta^{\alpha}(\sigma^{-2})^{\alpha-1}e^{-\beta\sigma^{-2}}}{\Gamma(\alpha)}\left|-2\sigma^{-3}\right| = \frac{2\beta^{\alpha}}{\Gamma(\alpha)}\sigma^{-2\alpha-1}e^{-\beta/\sigma^{2}}, \sigma > 0. \quad (1.7)$$

We find that:

- the prior mean of σ is

$$E\sigma = E\lambda^{-1/2} = \int_0^{\infty} \lambda^{-1/2}\frac{\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}d\lambda$$

$$= \frac{\beta^{\alpha}\Gamma(\alpha-1/2)}{\beta^{\alpha-1/2}\Gamma(\alpha)}\int_0^{\infty}\frac{\beta^{\alpha-\frac{1}{2}}\lambda^{\alpha-\frac{1}{2}-1}e^{-\beta\lambda}}{\Gamma(\alpha-1/2)}d\lambda$$

$$= \beta^{1/2}\frac{\Gamma(\alpha-1/2)}{\Gamma(\alpha)} = 0.9400$$

- the prior mode of σ is $Mode(\sigma) = \sqrt{\frac{2\beta}{2\alpha+1}} = 0.7559$

  (obtained by setting the derivative of the logarithm of (1.7) to zero, where that derivative is derived as follows:

$$l(\sigma) = \log f(\sigma) = -(2\alpha+1)\log\sigma - \beta\sigma^{-2} + \text{constant}$$

$$\Rightarrow l'(\sigma) = -\frac{2\alpha+1}{\sigma} + 2\beta\sigma^{-3} \overset{set}{=} 0 \Rightarrow \sigma^{2} = \frac{2\beta}{2\alpha+1})$$

- the prior median of σ is $Median(\sigma) = \sqrt{Median(\sigma^{2})} = 0.8648$
- the MLE of σ is $\hat{\sigma} = \sqrt{s_{\mu}^{2}} = 1.4754$ (which is biased).

By analogy with the above, $f(\sigma \mid y) = \frac{2b^{a}}{\Gamma(a)}\sigma^{-2a-1}e^{-b/\sigma^{2}}, \sigma > 0.$

So we find that:

- the posterior mean of σ is $E(\sigma \mid y) = b^{1/2}\frac{\Gamma(a-1/2)}{\Gamma(a)} = 1.1836$
- the posterior mode is $Mode(\sigma \mid y) = \sqrt{\frac{2b}{2a+1}} = 1.0262$

- the posterior median is

$$Median(\sigma \mid y) = \sqrt{Median(\sigma^2 \mid y)} = 1.1235$$

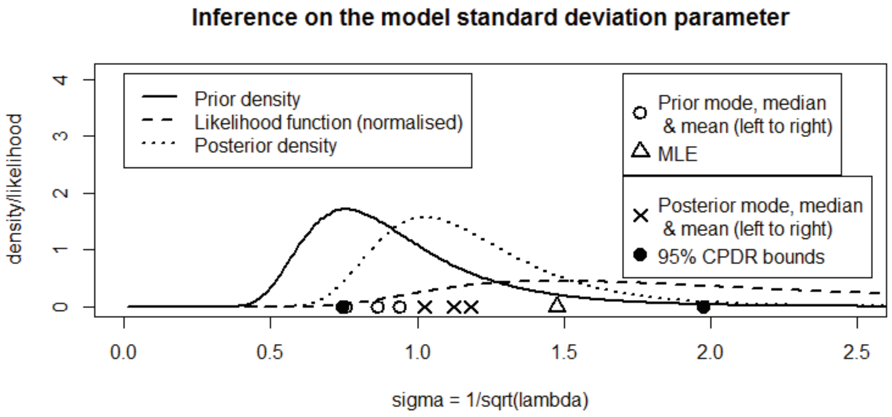
(since $1/2 = P(\sigma^2 < m \mid y) = P(\sigma < \sqrt{m} \mid y)$)

- the 95% CPDR for σ is $(0.7440, 1.9747)$ (where these bounds are the square roots of the bounds of the 95% CPDR for σ^2).

Figure 1.16 shows the various densities and estimates here, as well as the normalised likelihood function.

Note: The normalised likelihood function is the same as the posterior density of σ if the prior on σ is taken to be uniform over the positive real line, i.e. $f(\sigma) \propto 1, \sigma > 0$. This prior is specified by $\lambda \sim G(-1/2, 0)$, i.e. by $\alpha = -1/2$ and $\beta = 0$, as is evident from (1.7) above.

Figure 1.16 Results for Exercise 1.21(c)



(d) Considering the various point estimates of λ, σ^2 and σ derived above, we find that two of them can easily be expressed as credibility estimates, as follows. First, observe that

$$E(\sigma^2 \mid y) = \frac{b}{a-1} = \frac{\beta + ns_\mu^2/2}{\alpha + (n/2) - 1} = \frac{2\beta + ns_\mu^2}{2\alpha + n - 2}$$

$$= \left(\frac{n}{n + 2\alpha - 2}\right) s_\mu^2 + \frac{2\beta}{n + 2\alpha - 2},$$

where

$$\frac{2\beta}{n + 2\alpha - 2} = \frac{2\beta}{n + 2\alpha - 2} \times \frac{\alpha - 1}{\beta} \times \frac{\beta}{\alpha - 1} = \frac{2\alpha - 2}{n + 2\alpha - 2} \times E\sigma^2.$$

We see that the posterior *mean* of σ^2 is a credibility estimate of the form

$$E(\sigma^2 \mid y) = (1-c)E\sigma^2 + cs_\mu^2,$$

where:

$$E\sigma^2 = \frac{\beta}{\alpha-1} \text{ is the prior mean of } \sigma^2$$

$$s_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 \text{ is the MLE of } \sigma^2$$

$$c = \frac{n}{n+2\alpha-2} \text{ is the credibility factor (assigned to the MLE).}$$

Likewise,

$$Mode(\sigma^2 \mid y) = \frac{b}{a+1} = \frac{\beta + ns_\mu^2/2}{\alpha+(n/2)+1} = \frac{2\beta+ns_\mu^2}{2\alpha+n+2}$$

$$= \left(\frac{n}{n+2\alpha+2}\right)s_\mu^2 + \frac{2\beta}{n+2\alpha+2},$$

where

$$\frac{2\beta}{n+2\alpha+2} = \frac{2\beta}{n+2\alpha+2}\times\frac{\alpha+1}{\beta}\times\frac{\beta}{\alpha+1}$$

$$= \frac{2\alpha+2}{n+2\alpha+2}\times Mode(\sigma^2)$$

$$= \left(1-\frac{n}{n+2\alpha+2}\right)Mode(\sigma^2).$$

We see that the posterior *mode* of σ^2 is a credibility estimate of the form

$$Mode(\sigma^2 \mid y) = (1-d)Mode(\sigma^2) + ds_\mu^2,$$

where:

$$Mode(\sigma^2) = \frac{\beta}{\alpha+1} \text{ is the prior mode of } \sigma^2$$

$$s_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 \text{ is the MLE of } \sigma^2$$

(i.e. mode of the likelihood function)

$$d = \frac{n}{n+2\alpha+2} \text{ is the credibility factor (assigned to the MLE).}$$

R Code for Exercise 1.21

```
# (a) Inference on lambda ---------------------------------------------

y = c(8.4, 10.1, 9.4); n = length(y); mu=8;  alp=3; bet=2; options(digits=4)
a=alp+n/2; sigmu2=mean((y-mu)^2); b=bet+(n/2)*sigmu2

c(a,sigmu2,b) # 4.500 2.177 5.265

lampriormean=alp/bet; lamlikemode=1/sigmu2; lampriormode=(alp-1)/bet
lampriormedian= qgamma(0.5,alp,bet)
lampostmean=a/b; lampostmode=(a-1)/b; lampostmedian=qgamma(0.5,a,b)
lamcpdr=qgamma(c(0.025,0.975),a,b)

c(lampriormean,lamlikemode,lampriormode,lampriormedian,
          lampostmode,lampostmedian, lampostmean,lamcpdr)
  # 1.5000 0.4594 1.0000 1.3370 0.6648 0.7923 0.8547 0.2564 1.8065

lamv=seq(0,5,0.01); prior=dgamma(lamv,alp,bet)
post=dgamma(lamv,a,b);  like=dgamma(lamv,a-alp+1,b-bet+0)

X11(w=8,h=4); par(mfrow=c(1,1))

plot(c(0,5),c(0,1.9),type="n",
          main="Inference on the model precision parameter",
          xlab="lambda",ylab="density/likelihood")
lines(lamv,prior,lty=1,lwd=2); lines(lamv,like,lty=2,lwd=2);
lines(lamv,post,lty=3,lwd=2)
points(c(lampriormean,lampriormode, lampriormedian,
          lamlikemode,lampostmode,lampostmedian,lampostmean),
          rep(0,7),pch=c(1,1,1,2,4,4,4),cex=rep(1.5,7),lwd=2)
points(lamcpdr,c(0,0),pch=rep(16,2),cex=rep(1.5,2))

legend(0,1.9,
    c("Prior density","Likelihood function (normalised)","Posterior density"),
          lty=c(1,2,3),lwd=c(2,2,2))
legend(3,1.9,c("Prior mode, median\n & mean (left to right)",
          "MLE"), pch=c(1,2),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))
legend(3,1,c("Posterior mode, median\n & mean (left to right)",
          "95% CPDR bounds"), pch=c(4,16),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))
```

```
# (b)  Inference on sigma2 = 1/lambda ------------------------------------------------

sig2priormean=bet/(alp-1); sig2likemode=sigmu2; sig2priormode=bet/(alp+1)
sig2postmean=b/(a-1); sig2postmode=b/(a+1);
sig2postmedian=1/lampostmedian
sig2cpdr=1/qgamma(c(0.975,0.025),a,b); sig2priormedian= 1/lampriormedian

c(sig2priormean, sig2likemode, sig2priormode, sig2priormedian,
          sig2postmode, sig2postmedian, sig2postmean, sig2cpdr)
  # 1.0000 2.1767 0.5000 0.7479 0.9573 1.2622 1.5043 0.5535 3.8994

sig2v=seq(0.01,10,0.01); prior=dgamma(1/sig2v,alp,bet)/sig2v^2
post=dgamma(1/sig2v,a,b)/sig2v^2;
like=dgamma(1/sig2v,a-alp-1,b-bet+0)/sig2v^2

plot(c(0,10),c(0,1.2),type="n",
          main="Inference on the model variance parameter",
          xlab="sigma^2 = 1/lambda",ylab="density/likelihood")
lines(sig2v,prior,lty=1,lwd=2); lines(sig2v,like,lty=2,lwd=2)
lines(sig2v,post,lty=3,lwd=2)

points(c(sig2priormean, sig2priormode, sig2priormedian, sig2likemode,
          sig2postmode, sig2postmedian,sig2postmean),
          rep(0,7),pch=c(1,1,1,2,4,4,4),cex=rep(1.5,7),lwd=2)
points(sig2cpdr,c(0,0),pch=rep(16,2),cex=rep(1.5,2))

legend(1.8,1.2,
    c("Prior density","Likelihood function (normalised)","Posterior density"),
          lty=c(1,2,3),lwd=c(2,2,2))
legend(7,1.2,c("Prior mode, median\n & mean (left to right)",
          "MLE"), pch=c(1,2),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))
legend(6,0.65,c("Posterior mode, median\n & mean (left to right)",
          "95% CPDR bounds"), pch=c(4,16),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))

# abline(h=max(like),lty=3)  # Checking likelihood and MLE are consistent
# fun=function(t){  dgamma(1/t,a-alp-1,b-bet+0)/t^2  }
# integrate(f=fun,lower=0,upper=Inf)$value
             # 1  Checking likelihood is normalised
```

```
# (c)  Inference on sigma = 1/sqrt(lambda) --------------------------------------------

sigpriormean=sqrt(bet)*gamma(alp-1/2)/gamma(alp);
siglikemode=sqrt(sigmu2); sigpriormode=sqrt(2*bet/(2*alp+1))
sigpostmean= sqrt(b)*gamma(a-1/2)/gamma(a)
sigpostmode= sqrt(2*b/(2*a+1)); sigpostmedian=sqrt(sig2postmedian)
sigcpdr=sqrt(sig2cpdr); sigpriormedian= sqrt(sig2priormedian)

c(sigpriormean, siglikemode, sigpriormode, sigpriormedian,
          sigpostmode, sigpostmedian, sigpostmean, sigcpdr)
  # 0.9400 1.4754 0.7559 0.8648 1.0262 1.1235 1.1836 0.7440 1.9747

sigv=seq(0.01,3,0.01); prior=dgamma(1/sigv^2,alp,bet)*2/sigv^3
post=dgamma(1/sigv^2,a,b)*2/sigv^3;
like=dgamma(1/sigv^2,a-alp-1/2,b-bet+0)*2/sigv^3

plot(c(0,2.5),c(0,4.1),type="n",
          main="Inference on the model standard deviation parameter",
          xlab="sigma = 1/sqrt(lambda)",ylab="density/likelihood")
lines(sigv,prior,lty=1,lwd=2)
lines(sigv,like,lty=2,lwd=2)
lines(sigv,post,lty=3,lwd=2)
points(c(sigpriormean, sigpriormode, sigpriormedian, siglikemode,
          sigpostmode, sigpostmedian,sigpostmean),
          rep(0,7),pch=c(1,1,1,2,4,4,4),cex=rep(1.5,7),lwd=2)
points(sigcpdr,c(0,0),pch=rep(16,2),cex=rep(1.5,2))

legend(0,4.1,
    c("Prior density","Likelihood function (normalised)","Posterior density"),
          lty=c(1,2,3),lwd=c(2,2,2))
legend(1.7,4.1,c("Prior mode, median\n & mean (left to right)",
          "MLE"), pch=c(1,2),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))
legend(1.7,2.3,c("Posterior mode, median\n & mean (left to right)",
          "95% CPDR bounds"), pch=c(4,16),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))
```