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

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

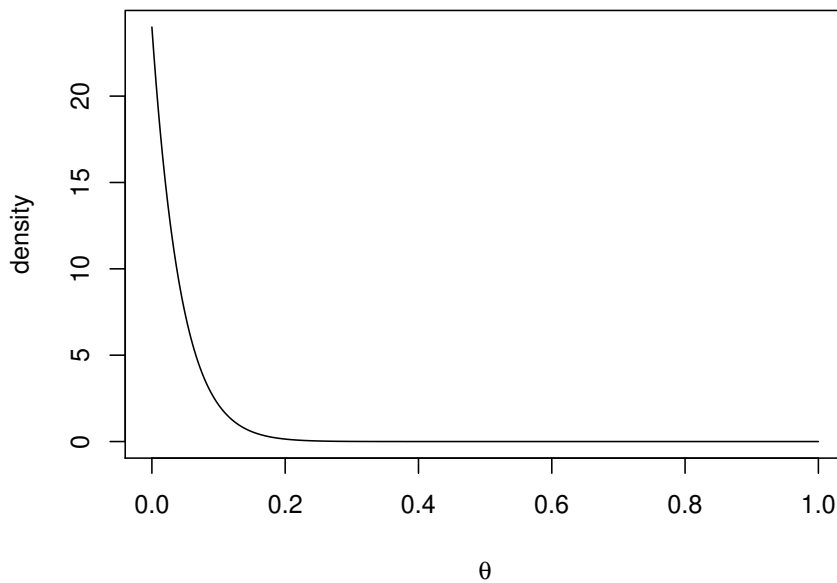A plot of this distribution is given in Figure 4.1. Determine the $100(1 - \alpha)\%$ HDI for θ.

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.
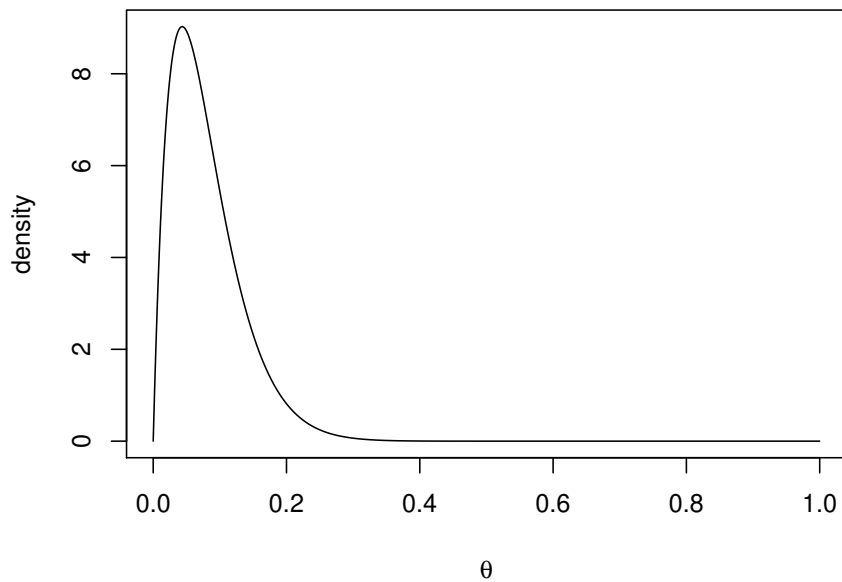

Figure 4.2: Plot of the $Beta(2, 23)$ posterior density function

✎...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

$$\pi(\theta = 1/2|x = 10) = \frac{\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2)}{[\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2) + \Pr(X = 10|\theta = 1/4)\pi(\theta = 1/4)} \\ + \Pr(X = 10|\theta = 1/8)\pi(\theta = 1/8)]$$

$$= \frac{0.1048 \times 0.2}{0.1048 \times 0.2 + 0.0413 \times 0.5 + 0.0008 \times 0.3}$$

$$= 0.501$$

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

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,

	correct	naive
y	$f(y\|x=1)$	$f(y\|\theta=\hat{\theta})$
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: $\mathrm{Pr}_{\pi^*}(\theta=1/23|x=1)=1$ 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 $y = 0, 1, 2, \ldots, 5$

$$f(y|x = 1) = \frac{f(y|\theta)\pi(\theta|x = 1)}{\pi(\theta|x = 1, y)}$$

$$= \frac{\binom{5}{y} \theta^y (1 - \theta)^{5-y} \times \dfrac{\theta(1 - \theta)^{22}}{B(2, 23)}}{\dfrac{\theta^{y+1}(1 - \theta)^{27-y}}{B(y + 2, 28 - y)}}$$

$$= \binom{5}{y} \frac{B(y + 2, 28 - y)}{B(2, 23)}.$$

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...*

...*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