Bayesian Inference

Lecturer: Loukia Meligkotsidou

February 17, 2017

Chapter 1

Introduction

1.1 What is statistical inference?

Before defining Bayesian inference we should consider the broader question, 'what is statistical inference?'. Many definitions are possible, but most boil down to the principle that statistical inference is the science of making conclusions about a 'population' from 'sample', items drawn from that population. This itself begs many questions about what is meant by a population, how the sample relates to the population, how we should enact the sampling if all options are available and so on. But we'll leave these issues aside, and focus our discussion on a simple example.

Suppose the Forestry Commission wish to estimate the proportion of trees in a large forest which suffer from a particular disease. It's impractical to check every tree, so they select a sample of just n trees. Again, we won't discuss here how they might choose their sample, but we'll suppose their sampling is random, in the sense that if θ is the proportion of trees having the disease in the forest, then each tree in the sample will have the disease, independently of all others in the sample, with probability θ. Denoting by X the random variable corresponding to the number of diseased trees in the sample, the Commission will use the observed value of $X = x$, to draw an inference about the population parameter θ. This inference could take the form of a *point estimate* ($\hat{\theta} = 0.1$); a *confidence interval* (95 % confident that θ lies in the range [0.08,0.12]); a *hypothesis test* (reject the hypothesis that $\theta < 0.07$ at the 5% significance level); or a *prediction* (predict that 15% of trees will be affected by next year).

In each case, knowledge of the observed sample value $X = x$ is being used to draw inferences about the population characteristic θ. Moreover, these inferences are made by specifying a probability model, $f(x|\theta)$, which determines how, for a given value of θ, the probabilities of the

different values of X are distributed.

Here for instance, under the assumptions made about random sampling, our model would be

$$X|\theta \sim \text{Binomial}\,(n, \theta)$$

Therefore

$$f(x \mid \theta) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$$

Statistical inference then amounts to an inference about the population parameter θ on the basis of observing $X = x$, and basically we'd infer that values of θ which give high probability to the value of x we observed are 'more likely' than those which assign x low probability — the principle of maximum likelihood. (Note that in its widest context, statistical inference also encompasses the issues of model choice, model verification etc., but we'll restrict attention to the inference of parameters within a parametric family of models.)

Before moving on to Bayesian inference in particular, there are some points to be made about this classical approach to inference. The most fundamental point is that the parameter θ, whilst not known, is being treated as *constant* rather than *random*. This is the cornerstone of classical theory, but leads to problems of interpretation. We'd like a 95% confidence interval of $[0.08, 0.12]$ to mean there's a 95% probability that θ lies between 0.08 and 0.12. It *cannot* mean this, since θ is not random: it either *is* in the interval, or it *isn't* — probability doesn't (and cannot) come in to it. The only random element in this probability model is the data, so the correct interpretation of the interval is that if we applied our procedure 'many times', then 'in the long run', the intervals we construct will contain θ on 95% of occasions. All inferences based on classical theory are forced to have this type of long–run–frequency interpretation, even though, for example, we only have the one interval $[0.08, 0.12]$ to interpret.

1.2 What is Bayesian inference?

The overall framework which Bayesian inference works within is identical to that above: there is a population parameter θ which we wish to make inferences about, and a probability mechanism $f(x \mid \theta)$ which determines the probability of observing different data x, under different parameter values θ. The fundamental difference however is that θ is treated as a *random* quantity. This might seem innocuous enough, but in fact leads to a substantially different approach to statistical modelling and inference.

In essence, our inference will be based on $f(\theta \mid x)$ rather than $f(x \mid \theta)$; that is the probability distribution of the parameter given the data, rather than the data given the parameter. In

many ways this leads to much more natural inferences, but to achieve this we will see that it is necessary to specify a *prior probability distribution*, $f(\theta)$, which represents beliefs about the distribution of θ *prior* to having any information about the data.

This notion of a prior distribution for the parameter θ is at the heart of Bayesian thinking, and depending on whether you are talking to an advocate or an opponent of the methodology, is either its primary advantage over classical theory or its biggest pitfall.

The coin sampling example.

This example is similar in idea to the forestry example, but sufficiently simple to illustrate the Bayesian approach to inference. In this example the use of the prior distribution is uncontroversial. We will note how the likelihood still has a central role in the Bayesian method.

Five coins have been placed on the table. You are required to estimate the proportion θ of these coins which are tails, by looking at a sample of just two coins. Now there are only 6 possible values of θ, written $M/5$ for $M = 0, 1, \ldots, 5$. Let X be the number of tails in the sample of two coins, and suppose we observe $X = 1$. In the *second* line of the table are the probabilities of observing this outcome, depending on the (unknown) value of θ.

To make sure you can work out these probabilities, consider for example $\mathrm{P}(X = 1|\theta = \frac{3}{5})$.

We have 3 tails and 2 heads in the set of 5 coins. We find 1 tail and 1 head in the sample of 2 coins.

The number of ways of picking these is $\binom{3}{1} \times \binom{2}{1} = 3 \times 2 = 6$, out of a total number of ways $\binom{5}{2} = 10$. The probability is then $\frac{6}{10} = 0.6$.

θ	$\frac{0}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{5}{5}$
$f(X = 1\|\theta)$	0.0	0.4	0.6	0.6	0.4	0.0
$f(\theta)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
$f(X = 1\|\theta) \times f(\theta) = f(X = 1, \theta)$	0	$\frac{2}{32}$	$\frac{6}{32}$	$\frac{6}{32}$	$\frac{2}{32}$	0
$f(\theta\|X = 1)$	0	$\frac{4}{32}$	$\frac{12}{32}$	$\frac{12}{32}$	$\frac{4}{32}$	0

Now the second row in the table is the *likelihood* of θ. It is *not* a probability distribution - it does not sum to 1. The most likely values of θ are $\frac{2}{5}$ and $\frac{3}{5}$; the values $\frac{0}{5}$ and $\frac{5}{5}$ are completely

excluded.

The Bayesian approach uses this likelihood function, but combines it with *prior* knowledge. The third line in the table shows the distribution of θ which we would plausibly assume even if we had *not* been allowed to look at a sample of 2 coins. Put another way, it is the distribution which would have plausibly described our knowledge *before* we had looked at the sample.

We use this by multiplying the second and third rows together to give part of the *joint* distribution of X and θ in the fourth row. This does not sum to 1 because it is only part of the joint distribution. In fact it sums to $\mathrm{P}(X = 1)$.

The important step is to get the conditional distribution of θ given $X = 1$ by dividing through by this sum. That is given in the fifth line. It *does* sum to 1; we made that happen by doing the division, a process we call *normalizing* a distribution.

This last row contains all the information we require to make any of the types of inference about θ that we have discussed. It is called the (Bayesian) posterior distribution of θ. It tells us that the values of $\theta = \frac{2}{5}$ and $\frac{3}{5}$ are equally probable, so it happens to give us a similar picture to the likelihood function in this case.

1.3 The prior distribution

In almost all situations, when we are trying to estimate a parameter θ, we do have some knowledge, or some belief, about the value of θ before we take account of the data. An example from O'Hagan (1994) makes this clear in qualitative terms.

You look out of your window and see a large wooden thing with branches covered by small green things. You entertain two possible hypotheses: one is that it's a tree, the other that it's the postman. Of course, you reject the postman hypothesis because postmen don't usually look like that, whereas trees do. Thus, in formal language, denoting by A the event that you see a wooden thing with green bits, B_1 the event it's a tree and B_2 the event it's the postman, you reject B_2 in favour of B_1 because $f(A|B_1) > f(A|B_2)$. Here, you're using the principle of maximising the likelihood.

But, you might also entertain a third possibility, B_3, that the thing is a replica of a tree. In this case it may well be that $f(A|B_1) = f(A|B_3)$, and yet you would still reject this hypothesis in favour of B_1. That is, even though the probability of seeing what you observed is the same whether it is a tree or a replica, your *prior* belief is that it's more likely to be a tree than a

replica and so you include this information when making your decision.

Consider another example, where in each of the following cases our data model is $X|\theta \sim Bin(10, \theta)$ and we observe $x = 10$ so that the hypothesis $H_0 : \theta \leq 0.5$ is rejected in favour of $H_1 : \theta > 0.5$ each time:

1. A woman tea–drinker claims she can detect from a cup of tea whether the milk was added before or after the tea. She does so correctly for ten cups.
2. A music expert claims she can distinguish between a page of Hayden's work and a page of Mozart. She correctly categorizes 10 pieces.
3. A drunk friend claims she can predict the outcome of tossing a fair coin, and does so correctly for 10 tosses.

Now, just in terms of the data, we would be forced to draw the same inferences in each case. But our prior beliefs suggest that we are likely to remain highly sceptical about the drunk friend, slightly impressed about the tea–drinker, and not surprised at all about the music expert.

The essential point is this: experiments are not abstract devices. Invariably, we have some knowledge about the process being investigated before obtaining the data. It is sensible (many would say essential) that inferences should be based on the combined information that this prior knowledge *and* the data represent. Bayesian inference is the mechanism for drawing inference from this combined knowledge.

Just to put the alternative point of view, it's this very reliance on prior beliefs which opponents of the Bayesian viewpoint object to. Different prior beliefs will lead to different inferences in the Bayesian view of things, and it's whether you see this as a good or a bad thing which determines your acceptability of the Bayesian framework.

1.4 Characteristics of the Bayesian approach

Following O'Hagan (1994), we can identify four fundamental aspects which characterize the Bayesian approach to statistical inference:

- **Prior Information.** All problems are unique and have their own context. That context derives prior information, and it is the formulation and exploitation of that prior knowledge which sets Bayesian inference apart from classical statistics.

- **Subjective Probability.** Classical statistics hinges on an objective 'long–run–frequency' definition of probabilities. Even if this is desirable, which is arguable, it leads to cumbersome inferences. By contrast, Bayesian statistics formalizes explicitly the notion that all probabilities are subjective, depending on an individual's beliefs and knowledge to hand. Thus, a Bayesian analysis is personalistic — unique to the specifications of each individual's prior beliefs. Inference is based on the *posterior* distribution $f(\theta|x)$, whose form will be seen to depend (through Bayes' theorem) on the particulars of the prior specification $f(\theta)$.

- **Self–consistency.** By treating the parameter θ as random, it emerges that the whole development of Bayesian inference stems quite naturally from probability theory only. This has many advantages, and means that all inferential issues can be addressed as probability statements about θ, which then derive directly from the posterior distribution. We will see one such advantage when we consider prediction in chapter 6.

- **No 'adhockery'.** Because classical inference cannot make probability statements about θ, various criteria are developed to judge whether a particular estimator is in some sense 'good'. This has led to a proliferation of procedures, often in conflict with one another. Bayesian inference sidesteps this tendency to invent *ad hoc* criteria for judging and comparing estimators by relying on the posterior distribution to express in straightforward probabilistic terms the entire inference about the unknown θ.

1.5 Objections to Bayesian inference

The main objections to Bayesian inference, as described above, are that the conclusions will depend on the specific choice of prior. This is particularly important for scientific papers, which should aim to be objective, and try and present the information in the data, not the beliefs of the author.

However, often there is prior information, that can sensibly be formulated in a prior distribution, and in such cases Bayesian inference is clearly the best approach.

Even when there is no prior information, it is possible to construct priors which reflect this lack of information (at least to some degree - see Chapter 3). Providing such uninformative priors are used, and there is sufficient data, then the posterior will closely resemble the likelihood, and the choice of prior will have had little effect on the inferences that are made.

1.6 Review of Bayes' Theorem

In its basic form, Bayes' Theorem is a simple result concerning conditional probabilities:

If A and B are two events with $\Pr(A) > 0$. Then

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

Proof.

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

The use of Bayes' Theorem, in probability applications, is to reverse the conditioning of events. That is, it shows how the probability of $B|A$ is related to $A|B$.

A slight extension of Bayes' Theorem is obtained by considering events $C_1, \ldots, C_k$ which partition the sample space Ω, so that $C_i \cap C_j = \phi$ if $i \neq j$ and $C_1 \cup \ldots \cup C_k = \Omega$. Then

$$\Pr(C_i|A) = \frac{\Pr(A|C_i)\Pr(C_i)}{\sum_{j=1}^{k} \Pr(A|C_j)\Pr(C_j)} \quad \text{for } i = 1, \ldots, k.$$

Proof.

$$\Pr(C_i|A) = \frac{\Pr(A \cap C_i)}{\Pr(A)} = \frac{\Pr(A|C_i)\Pr(C_i)}{\Pr(A)} \quad \text{for } i = 1, \ldots, k,$$

where

$$\Pr(A) = \sum_{i=1}^{k} \Pr(A \cap C_i) = \sum_{i=1}^{k} \Pr(A|C_i)\Pr(C_i),$$

since the C_i's are disjoint and their union is Ω.

A further extension is to continuous random variables:

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

Common Mistake 1 *When normalising a posterior distribution, take care with the limits on the parameters.*

Example 1.1 *A screening procedure for HIV is applied to a population which is at high risk for HIV; 10% of this population are believed to be HIV positive. The screening test is positive for 90% of people who are genuinely HIV positive, and negative for 85% of people who are not HIV positive. What are the probabilities of false positive and false negative results?*

Denoting A: person is HIV positive, and B: test result is positive, we have $\Pr(A) = 0.1$, $\Pr(B|A) = 0.9$ and $\Pr(B^c|A^c) = 0.85$.

We need to calculate Pr(false positive) $= \Pr(A^c|B)$ and Pr(false negative) $= \Pr(A|B^c)$.

We have:

$\Pr(A) = 0.1$ and $\Pr(A^c) = 1 - \Pr(A) = 0.9$

$\Pr(B \mid A) = 0.9$ and $\Pr(B^c \mid A) = 1 - \Pr(B \mid A) = 0.1$

$\Pr(B^c \mid A^c) = 0.85$ and $\Pr(B \mid A^c) = 1 - \Pr(B^c \mid A^c) = 0.15$

$\Pr(B) = \Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c) = 0.9 \times 0.1 + 0.15 \times 0.9 = 0.09 + 0.135 = 0.225$ and $\Pr(B^c) = 1 - \Pr(B) = 0.775$

So:

$$\Pr(A^c|B) = \frac{\Pr(B|A^c)\Pr(A^c)}{\Pr(B)} = \frac{0.15 \times 0.9}{0.225} = 0.6$$

and

$$\Pr(A|B^c) = \frac{\Pr(B^c|A)\Pr(A)}{\Pr(B^c)} = \frac{0.1 \times 0.1}{0.775} = 0.0129$$

Example 1.2 *In a bag there are 6 balls of unknown colours. Three balls are drawn without replacement and are found to be black. Find the probability that no black ball is left in the bag.*

So, let A: 3 black balls are drawn, and C_i: there were i black balls in the bag. Then, by Bayes' Theorem:

$$\Pr(C_i|A) = \frac{\Pr(A|C_i)\Pr(C_i)}{\sum_{j=0}^{6}\Pr(A|C_j)\Pr(C_j)}, \quad i = 0, \ldots, 6$$

But here's the key issue: what values do we give $\Pr(C_0), \ldots, \Pr(C_6)$? These are the probabilities of the different numbers of black balls in the bag, *prior* to having seen the data. Without any information to the contrary, we might well assume that all possible numbers are equally likely, so taking $\Pr(C_0) = \Pr(C_1) = \ldots = \Pr(C_6) = \frac{1}{7}$. In fact, we will use this prior specification for the remainder of the problem. But, is it the most sensible? You could take the view that it's quite likely that all balls in the bag are likely to be of the same colour, and consequently give higher prior probabilities to $\Pr(C_0)$ and $\Pr(C_7)$. Or you could find out from the ball manufacturers that they produce balls of 10 different colours. You might then take the prior view that each ball is black with probability $\frac{1}{10}$ and use this as a basis for calculating prior probabilities. The point is we have to *think hard* about how to express our prior beliefs, and the answer we get will depend on what we believe to begin with.

This argument will be taken up again later. For now though, using the prior specified above, we simply apply Bayes' Theorem to obtain:

$$\begin{aligned}
\Pr(C_3|A) &= \frac{\Pr(C_3)\Pr(A|C_3)}{\sum_{j=0}^{6}\Pr(A|C_j)\Pr(C_j)} \\
&= \frac{\frac{1}{7} \times (\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4})}{\frac{1}{7}\left\{0 + 0 + 0 + (\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4}) + (\frac{4}{6} \times \frac{3}{5} \times \frac{2}{4}) + (\frac{5}{6} \times \frac{4}{5} \times \frac{3}{4}) + (\frac{6}{6} \times \frac{5}{5} \times \frac{4}{4})\right\}} \\
&= \frac{1}{35}.
\end{aligned}$$

Thus, the data has updated our prior belief of $\Pr(C_3) = \frac{1}{7}$ to the posterior probability $\Pr(C_3|A) = \frac{1}{35}$. That is, the event is much less likely having seen the data than it was previously.

Stratum	Fossil present	Fossil absent
A	0.9	0.1
B	0.2	0.8

Table 1.1: Strata probabilities for fossil presence and absence.

1.7 Exercises

Exercise 1.1 *The rock strata A and B are difficult to distinguish in the field. Through careful laboratory studies it has been determined that the only characteristic which might be useful in aiding discrimination is the presence or absence of a particular brachipod fossil. In rock exposures of the size usually encountered, the probabilities of fossil presence are found to be as in table 1.1. It is also known that rock type A occurs about four times as often as type B in this area of study. If a sample is taken, and the fossil found to be present, calculate the posterior distribution of rock types.*

If the geologist always classifies as A when the fossil is found to be present, and classifies as B when it is absent, what is the probability she will be correct in a future classification?

Exercise 1.2 *Repeat Example 1.2 using a different choice of prior distribution. You may use any prior distribution but must give some convincing (but brief) reason for its use. In what way does this change of prior affect the posterior probability of no black balls left in the bag?*

Exercise 1.3 *Repeat the coin sampling example, by writing out the table for the case when the sample of 2 coins are observed both to be tails, i.e. $X = 2$. Just rows 2, 4 and 5 will change. In a word or two, say how the picture given by the posterior distribution differs from that of the likelihood in this case.*

Exercise 1.4 *A seed collector, who has acquired a small number of seeds from a plant, has a prior belief that the probability θ of germination of each seed is uniform over the range $0 \leq \theta \leq 1$. She experiments by sowing two seeds and finds that they both germinate. (i) Write down the likelihood function for θ deriving from this observation, and obtain the collector's posterior distribution of θ. (ii) Compute the posterior probability that θ is less than one half.*

Exercise 1.5 *A posterior distribution is calculated up to a normalising constant as*

$$f(\theta \mid x) \propto \theta^{-3},$$

for $\theta > 1$. Calculate the normalising constant of this posterior, and the posterior probability of $\theta < 2$.

Exercise 1.6 *A bee-keeper has 4 hives and in late winter his prior distribution of the number θ of hives in which the bees are still alive, is given in the table.*

θ	0	1	2	3	4
$f(\theta)$	0.3	0.1	0.1	0.2	0.3

He selects two hives at random and on careful inspection finds that the bees are alive in both.

(i) Write down the table of likelihoods of θ given this observation.

(ii) Evaluate the posterior probability that the bees are alive in all 4 of the hives.

Chapter 2

Bayesian updating

2.1 Introduction

As set out in Chapter 1, the essence of the Bayesian approach is to treat the unknown parameter θ as a random variable, specify a prior distribution for θ representing your beliefs about θ prior to having seen the data, use Bayes' Theorem to update prior beliefs into posterior probabilities, and draw appropriate inferences. Thus, there are four key steps to the Bayesian approach:

1. Specification of a likelihood model $f(x \mid \theta)$;
2. Determination of a prior $f(\theta)$;
3. Calculation of posterior distribution, $f(\theta \mid x)$ from Bayes' Theorem;
4. Drawing inferences from this posterior information.

In this chapter, we'll re–state Bayes' Theorem in a form appropriate for random variables rather than events, and consider some issues which arise when we try to use this result in the context of inference for a parameter θ. These issues will be addressed in subsequent chapters. We'll also look here at a number of examples where particular combinations of prior and likelihood give rise to convenient mathematical forms for the posterior distribution.

2.2 Bayes' Theorem

Stated in terms of random variables with densities denoted generically by f, Bayes Theorem takes the form:

$$f(\theta \mid x) = \frac{f(\theta)f(x \mid \theta)}{\int f(\theta)f(x \mid \theta)d\theta}$$

We will use this notation to cover the case where x is either continuous or discrete, where in the continuous case f is the p.d.f. as usual, but in the discrete case, f is the probability mass function of x. Similarly, θ can be discrete or continuous, but in the discrete case $\int f(\theta)f(x \mid \theta)d\theta$ is to be interpreted as $\sum_j f(\theta_j)f(x \mid \theta_j)$.

Notice that the denominator in Bayes' Theorem is a function of x only — θ having been 'integrated out'. Thus another way of writing Bayes' Theorem is

$$\begin{aligned} f(\theta \mid x) &= cf(\theta)f(x \mid \theta) \\ &\propto f(\theta)f(x \mid \theta) \\ &= h(\theta) \end{aligned}$$

or, in words, 'the posterior is proportional to the prior times the likelihood'. The constant of proportionality c, which may depend on x but not θ, is a normalising constant, determined so that the posterior distribution integrates to one.

There is a unique pdf, say $g(\theta)$ which is proportional to any given function $h(\theta)$, because $g(\theta)$ can be determined uniquely as $g(\theta) = ch(\theta)$ where $c = 1/\int h(\theta)d\theta$. We shall use this fact soon, to recognise standard pdfs which are proportional to certain posterior distributions.

This fact also allows us to remove any factors of $h(\theta)$, which in our application is $f(\theta)f(x|\theta)$, which do NOT depend upon θ, before carrying out the normalisation. Examples will illustrate this.

Common Mistake 2 *The posterior distribution is NOT $h(\theta)$; this is particularly important for making probability statements about θ (see Chapter 5) and for calculating predictive distributions (see Chapter 6).*

2.3 Issues

2.3.1 Choice of likelihood model

This depends on the mechanics of the problem to hand, and is the same problem faced using classical inference — what is the most suitable model for our data? Often, knowledge of the

structure by which the data is obtained may suggest appropriate models (Binomial sampling, or Poisson counts, for example), but often a model will be 'hypothesised' (Y is linearly related to X with independent Normal errors, for example) and its plausibility assessed later in the context of the data.

2.3.2 Choice of prior

This issue is fundamental to the Bayesian framework, and will be discussed at length in Chapter 3. However, some points should be noted now:

1. Because the prior represents your beliefs about θ before observing the data, it follows that the subsequent analysis is unique to you. Someone else's priors would lead to a different posterior analysis. In this sense the analysis is subjective.

2. We will see later that as long as the prior is not 'completely unreasonable' then the effect of the prior becomes less influential as more and more data become available. Thus, there is a sense in which mis–specification of the prior is unimportant so long as there is enough data available.

3. Often we might have a 'rough idea' what the prior should look like (perhaps we could give its mean and variance), but cannot be more precise than that. In such situations we could use a 'convenient' form for the prior which is consistent with our beliefs, but which also makes the mathematics relatively straightforward. We'll see some examples of this type of analysis shortly.

4. Sometimes we might feel that we have no prior information about a parameter. In such situations we might wish to use a prior which reflects our ignorance about the parameter. This is often possible, but there are some difficulties involved. These will be discussed in Chapter 3.

2.3.3 Computation

Though straightforward enough in principle, the implementation of Bayes' Theorem in practice can be computationally difficult, mainly as a result of the normalizing integral in the denominator. We will see that for some choices of prior–likelihood combination, this integral can be avoided, but in general, specialised techniques are required to simplify this calculation.

2.3.4 Inference

Bayesian analysis gives a more complete inference in the sense that all knowledge about θ available from the prior and the data is represented in the posterior distribution. That is, $f(\theta|x)$ *is* the inference. Still, it is often desirable to summarize that inference in the form of a point estimate, or an interval estimate. We will discuss this in Chapter 5. Moreover, desirable properties or concepts of statistics, functions of the data that are used for inferential purposes, are also present in Bayesian analysis. For example, in classical inference, the role of sufficient statistics is discussed in considerable detail. This concept has analogous role in Bayesian inference, but is more intuitively appealing. For instance, in Bayesian statistics, sufficiency can be characterised by saying that if we partition our data by $x = (x_1, x_2)$, then x_1 is sufficient for θ if $f(\theta|x)$ depends only on x_1 and does not depend on x_2.

2.4 Examples

Example 2.1 *(Binomial sample.) Suppose our likelihood model is* $X \sim \text{Binomial}\,(n, \theta)$*, and we wish to make inferences about* θ*, from a single observation* x*.*

So

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}; \qquad x = 0, \ldots, n.$$

Of course, in general, the choice of prior specification for θ will vary from problem to problem, and by definition will depend on the extent of our prior knowledge about the situation. However, we will proceed here by considering a possible family of prior distributions which, as we shall see, gives rise to simple computations. The point about this is that hopefully, provided the family is large enough and covers a sufficient range of possible forms, we can use a prior within this family which comes close to our true prior beliefs. If so, we get simple answers. If, however, there is no prior within this family which resembles what we really believe then we should avoid this approach.

So, in this case, suppose we can represent our prior beliefs about θ by a beta distribution:

$$\theta \sim \text{Beta}\,(p, q)$$

so that

$$\begin{aligned} f(\theta) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\theta^{p-1}(1-\theta)^{q-1} \qquad (0 \le \theta \le 1) \\ &\propto \theta^{p-1}(1-\theta)^{q-1}. \end{aligned}$$

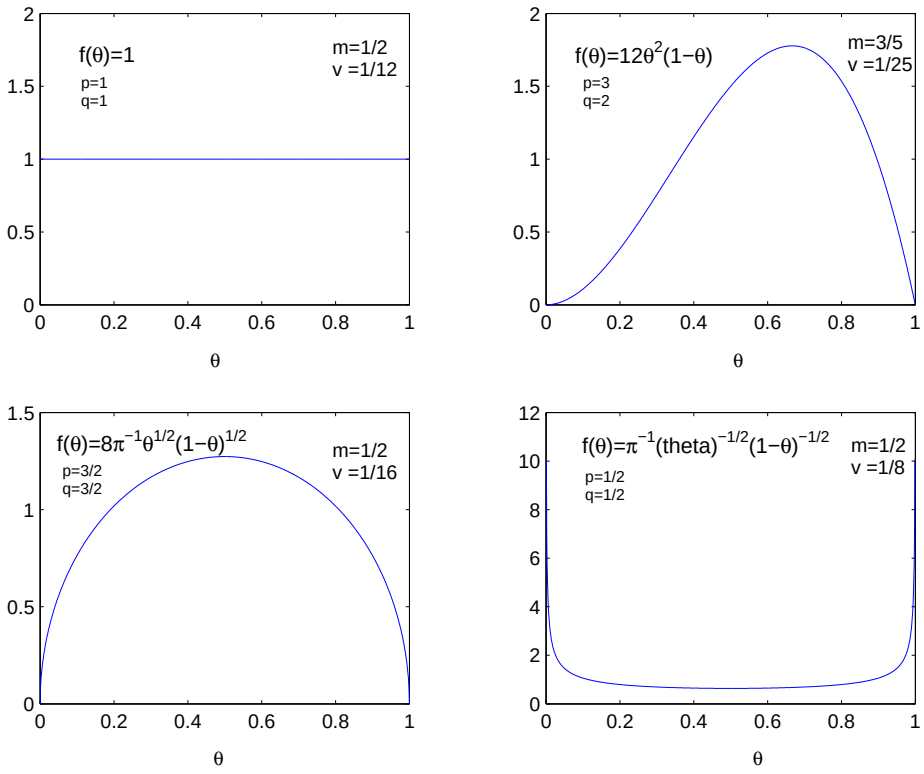

Figure 2.1: Cases of the beta distribution.

The parameters of this distribution are $p > 0$ and $q > 0$. (They are NOT probabilities and may have any positive value.) The mean and variance of this distribution are

$$\mathrm{E}(\theta) = m = \frac{p}{p+q} \quad \text{and} \quad \mathrm{Var}(\theta) = v = \frac{pq}{(p+q)^2(p+q+1)}. \tag{2.1}$$

The Beta distribution is also written

$$f(\theta) \;=\; \frac{\theta^{p-1}(1-\theta)^{q-1}}{\mathrm{B}(p,q)} \quad \text{where} \quad \mathrm{B}(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 \theta^{p-1}(1-\theta)^{q-1}\mathrm{d}\theta.$$

We call $\mathrm{B}(p,q)$ the beta function.

In Figure 2.1 are shown some simple cases of the beta distribution.

Now we apply Bayes' Theorem using this prior distribution:

$$\begin{aligned} f(\theta|x) &\propto f(\theta) f(x|\theta) \\ &\propto \theta^{p-1}(1-\theta)^{q-1} \times \theta^{x}(1-\theta)^{n-x} \\ &= \theta^{p+x-1}(1-\theta)^{q+n-x-1} \end{aligned}$$

$$= \quad \theta^{P-1}(1-\theta)^{Q-1}$$

where $P = p + x$ and $Q = q + n - x$. There is only one density function proportional to this, so it must be the case that

$$\theta|x \sim \text{Beta}\,(P, Q).$$

Thus, by careful choice, we have obtained a posterior distribution which is in the same family as the prior distribution, and in doing so have avoided the need to calculate explicitly any integrals for the normalising constant. That has been done already in the general formula for the beta distribution.

The effect of the data is to modify the parameters of the beta distribution from their prior values of (p, q), to the posterior values of $(p + x, q + n - x)$.

As a numerical example, consider the data set 'CANCER' taken from First Bayes.[1] Of 70 patients given a new treatment protocol for a particular form of cancer, 34 are found to survive beyond a specified period. Denote by θ the probability of a patient's survival. Consultation with medical experts, who are familiar with similar trials leads them to express the prior belief that $\text{E}\,(\theta) = 0.4$ and $\text{Var}\,(\theta) = 0.02$ (i.e. standard deviation about 0.14). Now, if a beta distribution is reasonable for their prior beliefs, then we should choose a prior distribution $\theta \sim \text{Beta}\,(p, q)$ such that $\text{E}\,(\theta) = 0.4$ and $\text{Var}\,(\theta) = 0.02$. Thus we require

$$m = \frac{p}{p+q} = 0.4 \qquad \text{and} \qquad v = \frac{pq}{(p+q)^2(p+q+1)} = 0.02.$$

These equations are solved by

$$p = \frac{(1-m)m^2}{v} - m \qquad \text{and} \qquad q = \frac{(1-m)^2 m}{v} - (1-m),$$

leading in this case to $p = 4.4$ and $q = 6.6$. This specifies the prior distribution for θ. This is a simple example of what is called *elicitation* of the prior distribution. In practice, it would now be necessary to ensure that the whole prior distribution was consistent with the experts' prior beliefs.

Presuming it is, we obtain $P = 4.4 + 34 = 38.4$ and $Q = 6.6 + 70 - 34 = 42.6$, so the posterior distribution of θ is

$$\theta|x \sim \text{Beta}\,(38.4, 42.6).$$

This posterior distribution summarizes all available information about θ and represents the complete inference about θ. We will discuss later how, if required, we might choose to summarize this inference, but for now we can see how the data modifies our prior beliefs by comparing prior

[1] First Bayes is a computer package, written by Tony O'Hagan, which illustrates a variety of Bayesian analyses.

and posterior expectations:

$$\mathrm{E}\,(\theta) = \frac{p}{p+q} = 0.4 \qquad\qquad \mathrm{E}\,(\theta|x) = \frac{P}{P+Q} = \frac{p+x}{p+q+n} = 0.474.$$

The effect of the observed data has been to increase the prior estimate of θ from 0.4 to 0.474. On the other hand, a natural estimate for θ on the basis of the data only is $x/n = 0.486$, which is the maximum likelihood estimate. Thus, the posterior estimate is a balance between our prior beliefs and the information provided by the data.

More generally, if x and n are large relative to p and q then the posterior expectation is approximately x/n, the maximum likelihood estimate. On the other hand, if p and q are moderately large then they will have reasonable influence on the posterior mean. It can also be checked that as x and n become larger — or indeed if p and q are chosen larger — then the posterior variance is lower.

Example 2.2 *(Poisson sample.) Suppose we have a random sample (i.e. independent observations) of size n, $x = (x_1, x_2, \ldots, x_n)$ of a random variable X whose distribution is* Poisson (θ) *, so that*

$$f(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad \theta \geq 0.$$

The mean and the variance of this distribution are:

$$E(X) = \theta$$

$$Var(X) = \theta$$

The likelihood is

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \\ &\propto e^{-n\theta}\theta^{\Sigma x_i} \end{aligned}$$

As in the binomial example, prior beliefs about θ will vary from problem to problem, but we'll look for a form which gives a range of different possibilities, but is also mathematically tractable.

In this case we suppose our prior beliefs can be represented by a gamma distribution:

$$\theta \sim \mathrm{Gamma}\,(p, q),$$

so

$$f(\theta) = \frac{q^p}{\Gamma(p)}\theta^{p-1}\exp\{-q\theta\}, \quad \theta > 0.$$

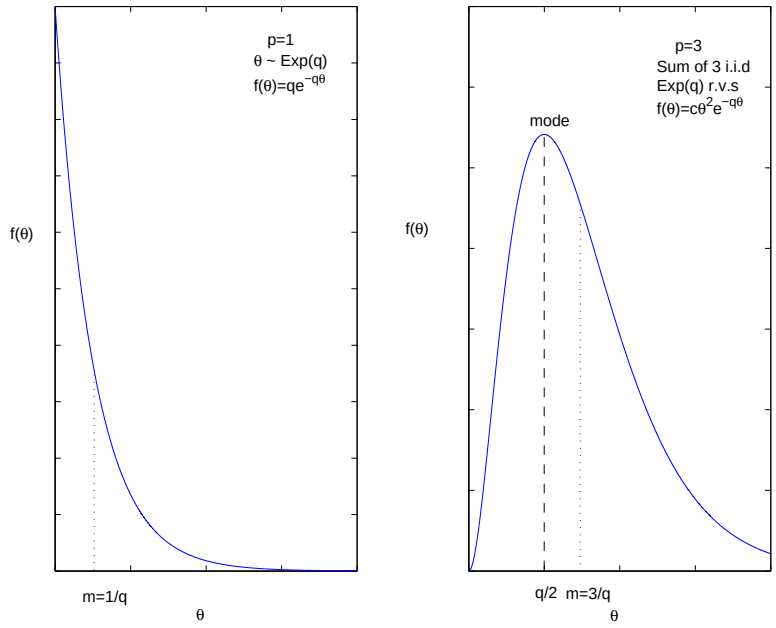

Figure 2.2: Examples of the gamma distribution.

The parameter $p > 0$ is a shape parameter, and $q > 0$ is a scale parameter. The mean and variance of this distribution are

$$\mathrm{E}(\theta) = m = \frac{p}{q} \quad \text{and} \quad \mathrm{Var}(\theta) = v = \frac{p}{q^2}. \tag{2.2}$$

In Figure 2.2 are shown some examples.

Applying Bayes' Theorem with this prior distribution,

$$\begin{aligned} f(\theta|x) &\propto \theta^{p-1}\exp\{-q\theta\} \times \exp\{-n\theta\}\theta^{\Sigma x_i} \\ &= \theta^{(p+\Sigma x_i - 1)}\exp\{-(q+n)\theta\} \\ &= \theta^{P-1}\exp(-Q\theta) \end{aligned}$$

where $P = p + \sum x_i$ and $Q = q + n$. Again, there is only one density function proportional to this, so it must be the case that

$$\theta|x \sim \mathrm{Gamma}(P, Q),$$

another gamma distribution whose parameters are modified by the sum of the data, $\sum_{i=1}^{n} x_i$, and the sample size n[2].

[2]Note that the individual values of x_i are not required, only their sum. We say that $\sum x_i$ is sufficient for θ.

Common Mistake 3 *When recognising the posterior distribution, remember that it is a distribution for the parameter, θ, not the data x_i. For example, it is a common mistake to recognise*

$$f(\theta \mid x) \propto \theta^{(p+\Sigma x_i - 1)} \exp\{-(q+n)\theta\},$$

as a Poisson distribution.

As a numerical example, again taken from First Bayes, we'll let θ be the mean number of geese in a flock within a particular region. Detailed aerial photographs of 45 flocks gave $\sum x_i = 4019$. We'll suppose our prior expectation and variance for θ are $p/q = m = 100$ and $p/q^2 = v = 20$ respectively. We can solve for $q = m/v = 5$ and $p = q \times m = 500$. Hence $P = 500 + 4019$ and $Q = 5 + 45$, so our posterior distribution is $\theta|x \sim \text{Gamma}\,(4519, 50)$. The mean and variance of this are:

$$m = \frac{4519}{50} = 90.4 \qquad v = \frac{90.4}{50} = 1.81$$

Example 2.3 *(Normal mean.) Let $x = (x_1, x_2, \ldots, x_n)$ be a random sample of size n of a random variable X with the* Normal (θ, σ^2) *distribution, where σ^2 is known:*

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

The mean and the variance of this distribution are:

$$E(X) = \theta$$

$$Var(X) = \sigma^2$$

Given a single observation x_i from the Normal (θ, σ^2) pdf, the likelihood of θ is given by

$$\begin{aligned} f(x_i \mid \theta) &= \tfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\tfrac{(x_i-\theta)^2}{2\sigma^2}\right\} \\ &\propto \exp\left\{-\tfrac{1}{2\sigma^2}(x_i^2 - 2x_i\theta + \theta^2)\right\} \\ &= \exp\left(-\tfrac{1}{2\sigma^2}x_i^2\right) \exp\left(\tfrac{1}{\sigma^2}x_i\theta - \tfrac{1}{2\sigma^2}\theta^2\right) \\ &\propto \exp\left(-\tfrac{1}{2\sigma^2}\theta^2 + \tfrac{1}{\sigma^2}x_i\theta\right). \end{aligned}$$

The likelihood of the whole sample is then

$$\begin{aligned} f(x \mid \theta) &= \textstyle\prod_i f(x_i \mid \theta) \\ &\propto \textstyle\prod_i \exp\left(-\tfrac{1}{2\sigma^2}\theta^2 + \tfrac{1}{\sigma^2}x_i\theta\right) \\ &= \exp\left[\textstyle\sum_i \left(-\tfrac{1}{2\sigma^2}\theta^2 + \tfrac{1}{\sigma^2}x_i\theta\right)\right] \\ &= \exp\left[-\left(\tfrac{n}{2\sigma^2}\right)\theta^2 + \left(\tfrac{1}{\sigma^2}\textstyle\sum_i x_i\right)\theta\right]. \end{aligned}$$

Suppose our prior beliefs about θ can themselves be represented as a normal distribution: $\theta \sim \text{Normal}\left(b, d^2\right)$. Again, this is to achieve simple mathematical analysis, but should only be used if such a choice *is* a good approximation to your prior belief about θ.

The mean and variance of this prior distribution are

$$\text{E}\left(\theta\right) = b \quad \text{and} \quad \text{Var}\left(\theta\right) = d^2. \tag{2.3}$$

Now, we can write our prior distribution for θ as

$$\begin{aligned} f(\theta) &= \tfrac{1}{\sqrt{2\pi}d} \exp\left\{-\tfrac{(\theta-b)^2}{2d^2}\right\} \\ &\propto \exp\left(-\tfrac{1}{2d^2}\theta^2 + \tfrac{1}{d^2}\, b\,\theta - \tfrac{1}{2d^2}b^2\right) \\ &\propto \exp\left(-\tfrac{1}{2d^2}\theta^2 + \tfrac{1}{d^2}\, b\,\theta\right). \end{aligned}$$

We now derive the posterior distribution of θ as follows

$$\begin{aligned} f(\theta \mid x) &\propto f(\theta) f(x \mid \theta) \\ &\propto \exp\left(-\tfrac{1}{2d^2}\,\theta^2 + \tfrac{1}{d^2} b\theta\right) \exp\left[-\left(\tfrac{n}{2\sigma^2}\right)\theta^2 + \left(\tfrac{1}{\sigma^2}\textstyle\sum_i x_i\right)\theta\right] \\ &= \exp\left[-\tfrac{1}{2}\left(\tfrac{1}{d^2} + \tfrac{n}{\sigma^2}\right)\theta^2 + \left(\tfrac{1}{d^2}b + \tfrac{1}{\sigma^2}\textstyle\sum_i x_i\right)\theta\right] \\ &= \exp\left(-\tfrac{1}{2D^2}\theta^2 + \tfrac{1}{D^2}B\theta\right). \end{aligned}$$

Therefore, we can conclude that the posterior distribution of θ is

$$\theta|x \;\sim\; \text{Normal}\left(B, D^2\right)$$

where

$$B = \text{E}\left(\theta|x\right) = \frac{\frac{1}{d^2}b + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{d^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad D^2 = \text{Var}\left(\theta|x\right) = \left(\frac{1}{d^2} + \frac{n}{\sigma^2}\right)^{-1},$$

and we have replaced $\sum x_i$ by $n\bar{x}$.

This result is expressed more concisely if we define 'precision' to be the reciprocal of variance: i.e. let $\tau = 1/\sigma^2$ and $c = 1/d^2$. Then

$$\theta|x \sim \text{Normal}\left(\tfrac{cb+n\tau\bar{x}}{c+n\tau}, \tfrac{1}{c+n\tau}\right)$$

Before looking at a numerical example, a number of observations can be made:

1. Observe that

$$E(\theta|x) = \gamma_n b + (1-\gamma_n)\overline{x}$$

where

$$\gamma_n = \frac{c}{c+n\tau} \quad \text{and} \quad (1-\gamma_n) = \frac{n\tau}{c+n\tau}.$$

Thus, the posterior mean is simply a weighted average of the prior mean and $\overline{x}$. Moreover, the weighting parameter γ_n is determined by the relative precision of the prior and data components. That is, if $n\tau$ is large relative to c, then $\gamma_n \approx 0$ and the posterior mean is close to $\overline{x}$.

2. Observe that 'posterior precision' = 'prior precision' + $n\times$ 'precision of each data item'.

3. As $n \to \infty$, then (loosely)

$$\theta|x \sim \text{Normal}\left(\overline{x}, \tfrac{\sigma^2}{n}\right)$$

so that the prior has no effect in the limit.

4. As $d \to \infty$, or equivalently $c \to 0$, we again obtain

$$\theta|x \sim \text{Normal}\left(\overline{x}, \tfrac{\sigma^2}{n}\right)$$

5. Note that the posterior distribution depends on the data only through $\overline{x}$ and not through the individual values of the x_i themselves. Again, we say that $\overline{x}$ is sufficient for θ.

Points 3 and 4 above are subtle, and will be discussed at some length below.

As a numerical example, taken from First Bayes, we look at a historical data set recorded by Henry Cavendish in the 18th century. He made 23 measurements of the density of the earth. For these data, $\overline{x} = 5.48$ and we'll suppose the variance of his measurement errors is known to be 0.04. Now, suppose from previous experiments the prior for θ, the earth's density, is taken as Normal $(5.4, 0.01)$. Then

$$B = \frac{\frac{1}{0.01}5.4 + \frac{23}{0.04}5.48}{\frac{1}{0.01} + \frac{23}{0.04}} \quad \text{and} \quad D^2 = \left(\frac{1}{0.01} + \frac{23}{0.04}\right)^{-1}$$

so $\theta|x \sim \text{Normal}\,(5.47, 0.00148)$.

2.5 General issues

The principles and details met in the above examples give rise to a number of issues which it is useful to discuss now.

2.5.1 Sequential updating

We have seen that Bayes' Theorem provides the machine by which your prior information is updated by data to give your posterior information. This then serves as your 'new' prior information before more data become available. This gives rise to one question in particular: if we obtain a sequence of data, and we update our beliefs on the arrival of each data item, would we get a different result from waiting till all the data had arrived, then updating our prior? Consider the simple case of two independent variables X_1 and X_2, each having density $f(x|\theta)$. Now, suppose we observe x_1, and update our prior via Bayes' Theorem to obtain

$$f(\theta|x_1) \propto f(\theta) \times f(x_1|\theta).$$

This becomes our new prior before observing x_2. Thus,

$$\begin{aligned} f(\theta|x_1, x_2) &\propto f(\theta) \times f(x_1|\theta) \times f(x_2|\theta) \\ &= f(\theta) \times f(x_1, x_2|\theta) \end{aligned}$$

which is the same result we would have obtained by updating on the basis of the entire information (x_1, x_2) directly. By induction this argument extends to sequences of any number of observations.

2.5.2 Sufficiency

In classical inference, sufficiency plays a central role both in theoretical development and in practical applications. The classical result by which we recognise that a function $s(x)$, of the data alone, is a sufficient statistic for a parameter θ, is that

$$f(x|\theta) = g(x)h(s, \theta)$$

where $g(x)$ does not involve θ, only the data (though not necessarily even the data - it could be some constant).

If this is the case, in the Bayesian analysis

$$f(x|\theta) \propto h(s, \theta)$$

so the likelihood depends on the data only through the sufficient statistic $s(x)$, as we have already seen a number of examples. In that case the posterior distribution $f(\theta|x)$ also depends on the data only through the sufficient statistic $s(x)$.

The reverse is also true, that if the posterior distribution $f(\theta|x)$ is seen to depend on the data only through a single function $s(x)$ of the data, then $s(x)$ is a sufficient statistic for θ.

2.5.3 The Likelihood Principle

The likelihood principle states that if two experiments yield the same likelihood (up to proportionality), then the inference we draw about θ should be the same in each case. In other words, all aspects of inference should be based only on the likelihood function. A major virtue of the Bayesian framework is that Bayesian techniques are inherently consistent with the likelihood principle, whereas many simple procedures from classical statistics violate it.

Example 2.4 *Consider two experiments concerned with estimating the probability of success θ in independent trials. In the first experiment, the number x of successes in n trials is recorded. In the second, the number y of trials required to obtain m successes is recorded.*

The distributions of the random variables X, Y describing the outcomes of these experiments differ. They are the Binomial and Negative Binomial distributions.

$$f(x|\theta) = P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n$$

and

$$f(y|\theta) = P(Y = y) = \binom{y-1}{m-1}\theta^m(1-\theta)^{y-m}, \quad y = m, m+1, \ldots, \infty.$$

The corresponding maximum likelihood estimators are $\hat{\theta} = x/n$ and $\hat{\theta} = m/y$. Their sampling distributions are quite different. If $n = 2$, then x/n can take the values 0, $\frac{1}{2}$ and 1. If $m = 1$, then m/y can take the values 1, $\frac{1}{2}$, $\frac{1}{3}$, ... But if it happened that also $x = 1$ and $y = 2$,

$$f(x|\theta) = 2\theta(1-\theta) \quad \text{and} \quad f(y|\theta) = \theta(1-\theta)$$

so that the likelihoods are proportional, and the Bayesian inference would be the same in both cases. *It would not depend on the probability of any other possible outcome of the experiment, i.e. not on the distribution of X or Y.*

2.6 Exercises

Exercise 2.1 *In each of the following cases, derive the posterior distribution:*
a) $x_1, \ldots, x_n$ *are a random sample from the distribution with probability function*

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

with the $\text{Beta}\,(p, q)$ *prior distribution*

$$f(\theta) = \frac{\theta^{p-1}(1-\theta)^{q-1}}{\text{B}(p, q)}, \quad 0 \leq \theta \leq 1.$$

b) $x_1, \ldots, x_n$ *are a random sample from the distribution with probability density function*

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!} \quad x = 0, 1, \ldots$$

with the prior distribution

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

Exercise 2.2 *The proportion,* θ*, of defective items in a large shipment is unknown, but expert assessment assigns* θ *the* $\text{Beta}\,(2, 200)$ *prior distribution. If 100 items are selected at random from the shipment, and 3 are found to be defective, what is the posterior distribution of* θ*?*

If another statistician, having observed the 3 defectives, calculated her posterior distribution as being a beta distribution with mean 4/102 and variance 0.0003658, then what prior distribution had she used?

Exercise 2.3 *The diameter of a component from a long production run varies according to a* $\text{Normal}\,(\theta, 1)$ *distribution. An engineer specifies that the prior distribution for* θ *is* $\text{Normal}\,(10, 0.25)$*. In one production run 12 components are sampled and found to have a sample mean diameter of 31/3. Use this information to find the posterior distribution of mean component diameter. Hence calculate the probability that this is more than 10 units.*

Exercise 2.4 *The number of defects in a single roll of magnetic tape has a* $\text{Poisson}\,(\theta)$ *distribution. The prior distribution for* θ *is* $\text{Gamma}\,(3, 1)$*. When 5 rolls of this tape are selected at random, the number of defects found on each are 2,2,6,0 and 3 respectively. Determine the posterior distribution of* θ*.*

Exercise 2.5 *Suppose that the time* x *in minutes required to serve a customer in a bank has an* $\text{Exp}\,(\theta)$ *distribution with pdf* $\theta \exp(-\theta x)$*, for* $x \geq 0$*. A prior for* θ *is determined to be a gamma*

distribution with mean 0.2 and standard deviation 1. The times x_1, x_2, ..., x_{20} taken to serve 20 customers are observed, and their average value $\bar{x}$ is 3.8 minutes. Determine the posterior distribution for θ.

Exercise 2.6 *A random sample $x_1, \ldots, x_n$ is taken from a Poisson distribution with mean θ. The prior distribution for θ is a gamma distribution with mean μ_0. If the sample mean is $\overline{x}_n$, show that the mean of the posterior distribution of θ will be a weighted average of the form*

$$\gamma_n \overline{x}_n + (1 - \gamma_n)\mu_0,$$

and show that $\gamma_n \to 1$ as $n \to \infty$.

Exercise 2.7 *An engineer uses a* Gamma (p, q) *distribution to describe the prior distribution of a parameter θ, the expected number of faults x to be found in a plastic moulding.*

(a) Given that the mean and variance of the gamma distribution are p/q and p/q^2 respectively, what values should she take for p and q in order that θ should have mean 12 and standard deviation 4.

(b) The engineer then ascertains that there are a total of 37 faults in 5 of the plastic mouldings. Assuming that the numbers of faults in each moulding are independent and have a Poisson distribution, obtain the posterior distribution of θ given this information.

(c) What is the posterior mean of θ?

Exercise 2.8 *(i) Observations $y_1, y_2, \ldots, y_n$ are obtained from independent random variables which are normally distributed, each with the same (known) variance σ^2 but with respective means $x_1\theta, x_2\theta, \ldots, x_n\theta$. The values of $x_1, x_2, \ldots, x_n$ are known but θ is unknown. Show that the likelihood, given a single observation y_i, is of the form*

$$f(y_i|\theta) \propto \exp\left(-\frac{1}{2}\frac{1}{\sigma^2}x_i^2\theta^2 + \frac{1}{\sigma^2}y_i x_i \theta\right).$$

(ii) Given the prior distribution for the unknown coefficient θ may be described as normal with mean b and variance σ^2/a^2, show that the posterior distribution of θ is proportional to

$$\exp\left\{-\frac{1}{2}\left[(a^2 + \sum_{i=1}^{n} x_i^2)/\sigma^2\right]\theta^2 + \left[(a^2 b + \sum_{i=1}^{n} y_i x_i)/\sigma^2\right]\theta\right\}.$$

(iii) Use this to write down the posterior mean of θ. Show that it may be written as

$$\hat{\theta} = wb + (1 - w)\frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2},$$

and obtain an expression for w.

Chapter 3

Specifying priors

3.1 Introduction

We have seen that the fundamental difference between Bayesian and classical statistics is that in Bayesian statistics unknown parameters are treated as random variables, and that the use of Bayes' theorem requires the specification of prior distributions for these parameters. Whilst this facilitates the inclusion of genuine prior belief about parameters, the choice of prior distribution cannot be made blindly; considerable care is needed and there are some very substantial issues involved. In this chapter we look at some of these issues.

3.2 Conjugate priors

The computational difficulties arise in using Bayes' Theorem when it is necessary to evaluate the normalizing constant in the denominator,

$$\int f(\theta) f(x|\theta) d\theta.$$

For example, suppose $X_1, \ldots, X_n$ are independent Poisson(θ) variables, and our beliefs about θ are that it *definitely* lies in the range $[0, 1]$, but that all values within that range are equally likely: thus, $f(\theta) = 1;\ 0 \leq \theta \leq 1$ and $f(\theta|x) \propto \exp(-n\theta)\theta^{\Sigma x_i}$. Then the normalizing constant is

$$\int_0^1 \exp(-n\theta)\theta^{\Sigma x_i} d\theta,$$

and this integral, an incomplete gamma function, can only be evaluated numerically.

So, even simple choices of priors can lead to awkward numerical problems. However, we saw three examples in the previous chapter where judicious choices of prior led to posterior calculations which did not require any integration. In each of these cases we were able to identify a prior distribution for which the posterior distribution was in the same family of distributions as the prior; such priors are called *conjugate priors*. Let's consider another example where a conjugate prior can be found.

Example 3.1 *(Gamma sample.) Let $X_1, \ldots X_n$ be independent variables having the* Gamma (k, θ) *distribution, where k is known. Note that the case $k = 1$ corresponds to the exponential distribution.*

Then

$$\begin{aligned} f(x_i \mid \theta) &= \frac{1}{\Gamma(k)} \theta^k x_i^{k-1} e^{-\theta x_i} \\ &\propto \theta^k e^{-\theta x_i} \end{aligned}$$

So

$$\begin{aligned} f(x \mid \theta) &\propto \prod_{i=1}^{n} \theta^k e^{-\theta x_i} \\ &= \theta^{nk} \exp\{-\theta \Sigma x_i\}. \end{aligned}$$

Now, studying this form, regarded as a function of θ suggests we could take a prior of the form

$$f(\theta) \propto \theta^{p-1} \exp\{-q\theta\}$$

that is, $\theta \sim$ Gamma (p, q), since then by Bayes' Theorem

$$f(\theta|x) \propto \theta^{p+nk-1} \exp\{-(q + \Sigma x_i)\theta\},$$

and so $\theta|x \sim$ Gamma $(p + nk, q + \sum x_i)$.

3.2.1 Use of conjugate priors

The use of conjugate priors should be seen for what it is: a convenient mathematical device. However, expression of one's prior beliefs as a parametric distribution is always an approximation. In many situations the richness of the conjugate family is great enough for a conjugate prior to be found which is sufficiently close to one's beliefs for this extra level of approximation to be acceptable. However, if this is not the case, they should not be used just because they made the maths easier.

3.2.2 Obtaining conjugate priors

Provided they are not in direct conflict with our prior beliefs, and provided such a family can be found, the simplicity induced by using a conjugate prior is compelling. But in what situations can a conjugate family be obtained?

It emerges that the only case where conjugates can be easily obtained is for data models within the *exponential family*. That is,

$$f(x|\theta) = h(x)g(\theta)\exp\{t(x)c(\theta)\}$$

for functions h, g, t and c such that

$$\int f(x|\theta)dx = g(\theta)\int h(x)\exp\{t(x)c(\theta)\}dx = 1.$$

This might seem restrictive, but in fact includes the exponential distribution, the Poisson distribution, the gamma distribution with known shape parameter, the binomial distribution and the normal distribution with known variance.

Given a random sample $x = (x_1, x_2, \ldots, x_n)$ from this general distribution, the likelihood for θ is then

$$\begin{aligned} f(x \mid \theta) &= \prod_{i=1}^{n}\{h(x_i)\}g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i)c(\theta)\} \\ &\propto g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i)c(\theta)\}. \end{aligned}$$

Thus if we choose a prior of the form

$$f(\theta) \propto g(\theta)^d \exp\{b\, c(\theta)\},$$

we obtain

$$\begin{aligned} f(\theta|x) &\propto f(\theta)f(x \mid \theta) \\ &\propto g(\theta)^d \exp\{b\, c(\theta)\} \times g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i)c(\theta)\} \\ &= g(\theta)^{n+d} \exp\{[b + \sum_{i=1}^{n} t(x_i)]c(\theta)\} \\ &= g(\theta)^D \exp\{Bc(\theta)\}, \end{aligned}$$

where $D = n + d$ and $B = b + \sum t(x_i)$. This results in a posterior in the same family as the prior, but with modified parameters.

It is easily checked that all the examples of conjugate priors we have met so far can be derived in this way. For example, a binomial random variable has pdf

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x \\ &= \binom{n}{x}(1-\theta)^n\exp\{x\log(\frac{\theta}{1-\theta})\}. \end{aligned}$$

So, in exponential family notation we have $h(x) = \binom{n}{x}$, $g(\theta) = (1-\theta)^n$, $t(x) = x$, and $c(\theta) = \log(\frac{\theta}{1-\theta})$. Thus, we construct a conjugate prior with the form

$$\begin{aligned} f(\theta) &\propto [(1-\theta)^n]^d\exp\{b\log(\frac{\theta}{1-\theta})\} \\ &= (1-\theta)^{nd-b}\theta^b \end{aligned}$$

which is a member of the beta family of distributions.

3.2.3 Standard conjugate analyses

Table 3.1 lists many of the standard prior–likelihood conjugate analyses.

Likelihood	Prior	Posterior
$x \sim \text{Binomial}\,(n,\theta)$	$\text{Beta}\,(p,q)$	$\text{Beta}\,(p+x, q+n-x)$
$x_1,\ldots,x_n \sim \text{Geometric}\,(\theta)$	$\text{Beta}\,(p,q)$	$\text{Beta}\,(p+n, q+\sum_{i=1}^n x_i - n)$
$x \sim \text{Negative-Binomial}\,(n,\theta)$	$\text{Beta}\,(p,q)$	$\text{Beta}\,(p+n, q+x-n)$
$x_1,\ldots,x_n \sim \text{Poisson}\,(\theta)$	$\text{Gamma}\,(p,q)$	$\text{Gamma}\,(p+\sum_{i=1}^n x_i, q+n)$
$x_1,\ldots,x_n \sim \text{Gamma}\,(k,\theta)$ (k known)	$\text{Gamma}\,(p,q)$	$\text{Gamma}\,(p+nk, q+\sum_{i=1}^n x_i)$
$x_1,\ldots,x_n \sim \text{Normal}\,(\theta,\tau^{-1})$, ($\tau$ known)	$\text{Normal}\,(b,c^{-1})$	$\text{Normal}\left(\frac{cb+n\tau\overline{x}}{c+n\tau}, \frac{1}{c+n\tau}\right)$

Table 3.1: Standard conjugate analyses.

Some of these have been given as examples; the others you should verify.

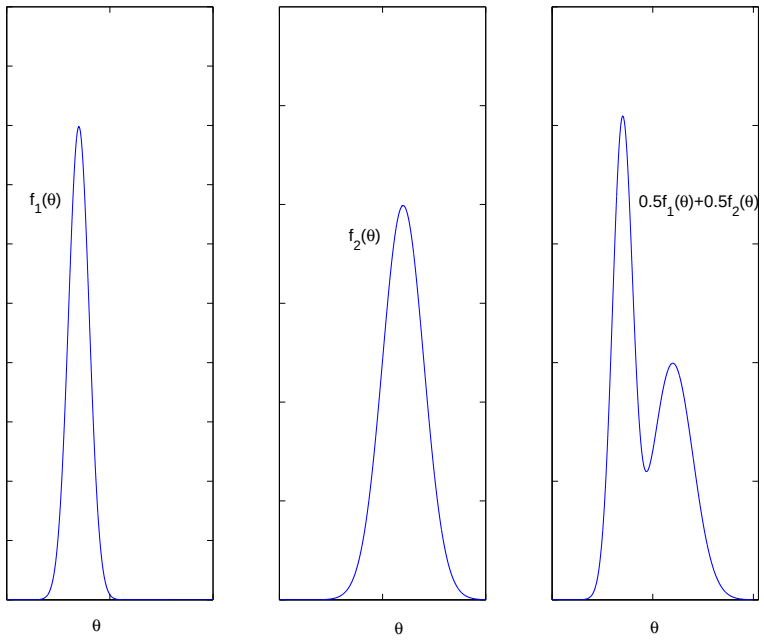

Figure 3.1: A mixture of prior distributions.

3.3 Mixtures of priors

The arguments for the use of conjugate families of prior distributions were set out in Section 3.1. However, we emphasise again, that such families should only be used if a suitable member of the family can be found which is in accord with your true prior beliefs. In some situations the natural conjugate family may be too restrictive for this to be possible.

Consider the following example (which is often quoted). When a coin is tossed, then almost invariably there is a 0.5 chance of it coming up heads. However, if the coin is *spun* on a table, it is often the case that slight imperfections in the edge of the coin cause it to have a tendency to prefer either heads or tails. Taking this into account, we may wish to give the probability θ of the coin coming up heads a prior distribution which favours values around either 0.3 or 0.7 say. That is, our prior beliefs may be reasonably represented by a *bimodal* distribution (or even *trimodal* if we wish to give extra weight to the unbiased possibility, $\theta = 0.5$). Our likelihood model for the number of heads in n spins will be binomial: $X|\theta \sim \text{Binomial}\,(n, \theta)$ and so the conjugate prior family is the beta family. However, no member of this family is multimodal. One solution is to use *mixtures* of conjugate distributions (see Figure 3.1).

This extended family will also be a conjugate prior family for the following reason. Suppose $f_1(\theta), \ldots, f_k(\theta)$ are all conjugate distributions for θ, leading to posterior distributions $f_1(\theta|x), \ldots, f_k(\theta|x)$. Now consider the family of mixture distributions:

$$f(\theta) = \sum_{i=1}^{k} p_i f_i(\theta),$$

where $0 \leq p_i \leq 1$, $i = 1, \ldots, k$ and $\sum_{i=1}^{k} p_i = 1$. Then,

$$\begin{aligned} f(\theta|x) &\propto f(\theta) f(x|\theta) \\ &= \sum_{i=1}^{k} p_i f_i(\theta) f(x|\theta) \\ &= \sum_{i=1}^{k} p_i f_i(x) f_i(\theta|x) \end{aligned}$$

Hence

$$\sum_{i=1}^{k} p_i^* f_i(\theta|x),$$

where $p_i^* \propto p_i f_i(x)$. So the posterior is in the same mixture–family. Notice though that the mixture proportions in the posterior p_i^* generally will be different from those in the prior.

It can be proved that finite mixtures of conjugate priors can be made arbitrarily close to *any* prior distribution. However, the number of terms in the mixing may be large, and it may be possible to represent one's prior beliefs much more succinctly using other non–conjugate families of models.

3.4 Improper priors

Consider again the posterior analysis obtained when estimating a normal mean with known variance, and using a normal prior. Thus, $X_1, \ldots, X_n \sim \text{Normal}\left(\theta, \tau^{-1}\right)$, $\theta \sim \text{Normal}\left(b, c^{-1}\right)$, leading to $\theta|x \sim \text{Normal}\left(\frac{cb+n\tau\overline{x}}{c+n\tau}, \frac{1}{c+n\tau}\right)$. Now, the strength of our prior beliefs about θ are determined by the variance, or equivalently the precision, c, of the normal prior. A large value of c corresponds to very strong prior beliefs; on the other hand small values of c reflect very weak prior information. Now, suppose our prior beliefs about θ were so weak that we let $c \to 0$. Then simply enough, the posterior distribution becomes Normal $\left(\overline{x}, \frac{1}{n\tau}\right)$, or in the more familiar notation: Normal $\left(\overline{x}, \frac{\sigma^2}{n}\right)$. Thus we seemingly obtain a perfectly valid posterior distribution through this limiting procedure.

But there's a catch. Consider what's happening to the prior as $c \to 0$. In effect, we obtain a Normal (b, ∞) distribution, which is not a genuine distribution. In fact, as $c \to 0$, the distribution

of Normal $\left(b, c^{-1}\right)$ becomes increasingly flatter, so that in any interval $-K \leq \theta \leq K$, provided c is sufficiently close to 0, we have approximately

$$f(\theta) \propto 1; \quad -K \leq \theta \leq K.$$

But this cannot be valid, in the limit as $c \to 0$, over the whole real line $\mathcal{R}$, because

$$\int_{\mathcal{R}} f(\theta) d\theta = \infty.$$

So, the posterior Normal $\left(\overline{x}, \frac{\sigma^2}{n}\right)$, obtained by letting $c \to 0$ in the standard conjugate analysis, cannot arise through the use of any proper prior distribution. It does arise however by formal use of the prior specification $f(\theta) \propto 1$, which is an example of what is termed an *improper* prior distribution.

So, is it valid to use a posterior distribution obtained by specifying an improper prior to reflect vague knowledge? Although there are some further difficulties involved (see below), generally the use of improper prior distributions is considered to be acceptable. The point really is that if we chose c to be any value other than zero, we would have obtained a perfectly proper prior and there would have been no qualms about the subsequent analysis. Thus, we could choose c arbitrarily close to zero and obtain a posterior arbitrarily close to the one we actually obtained by using the improper prior $f(\theta) \propto 1$.

3.5 Representations of ignorance

In the previous section, we saw that attempting to represent ignorance within the standard conjugate analysis of a Normal mean led to the concept of improper priors. But there are more fundamental problems as well. Consider that we might have specified a prior $f_\Theta(\theta)$ for a parameter θ in a model. It is quite reasonable to decide to use instead the parameter $\phi = 1/\theta$. For example θ may be the parameter of the exponential distribution of inter-arrival times in a queue, which represents the arrival rate. Then ϕ represents the mean inter-arrival time. By probability theory the corresponding prior density for ϕ must be given by

$$\begin{aligned} f_\Phi(\phi) &= f_\Theta(\theta) \times \left|\frac{d\theta}{d\phi}\right| \\ &= f_\Theta(1/\phi)\frac{1}{\phi^2}. \end{aligned}$$

In Figure 3.2 are shown plots of the prior distributions of a parameter θ and of $\phi = 1/\theta$.

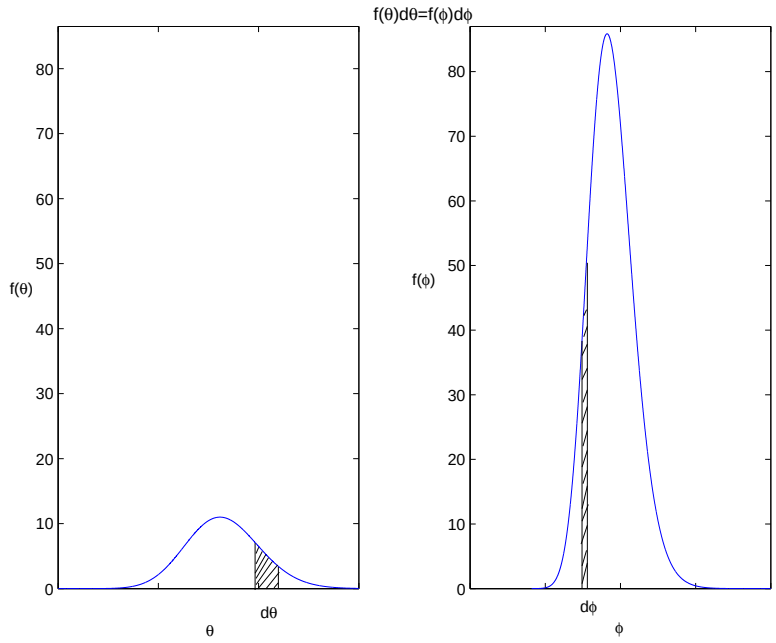

Figure 3.2: Jeffreys' priors of θ and $\phi = 1/\theta$.

If we decided that we wished to express our ignorance about θ by choosing $f_\Theta(\theta) \propto 1$, then we are forced to take $f_\Phi(\phi) \propto 1/\phi^2$. But if we are ignorant about θ, we are surely equally ignorant about ϕ, and so might equally have made the specification $f_\Phi(\phi) \propto 1$. Thus, prior ignorance as represented by uniformity of belief, is not preserved under re-parameterisation.

There is one way of using the likelihood $f(x|\theta)$, or more accurately, the log likelihood $\ell(\theta) = \log f(x \mid \theta)$, to specify a prior which *is* consistent across 1—1 parameter transformations. This is the 'Jeffreys' prior', and is based on the concept of Fisher information:

$$I(\theta) = -\mathrm{E}\left\{\frac{d^2\,\ell(\theta)}{d\theta^2}\right\} = \mathrm{E}\left\{\left(\frac{d\,\ell(\theta)}{d\theta}\right)^2\right\}.$$

Then, the Jeffreys' prior is defined as

$$J_\Theta(\theta) \propto |I(\theta)|^{\frac{1}{2}}.$$

The consistency of this prior under a parameter transformation $\phi(\theta)$ may be stated as:

$$J_\Phi(\phi) = J_\Theta(\theta)\left|\frac{d\theta}{d\phi}\right|.$$

Substituting the definition of Jeffrey's prior's, and squaring, we need to verify that

$$I(\phi) = I(\theta)\left|\frac{d\theta}{d\phi}\right|^2.$$

We can do this by using the chain rule for change of variable in differentiation. We have

$$\ell_\Phi(\phi) = \ell_\Theta(\theta(\phi)).$$

and

$$\frac{d\ell_\Phi(\phi)}{d\phi} = \frac{d\ell_\Theta(\theta)}{d\theta}\frac{d\theta(\phi)}{d\phi}.$$

Therefore

$$I(\phi) = \mathrm{E}\left\{\left(\frac{d\,\ell(\phi)}{d\phi}\right)^2\right\} = \mathrm{E}\left\{\left(\frac{d\,\ell(\theta)}{d\theta}\frac{d\theta}{d\phi}\right)^2\right\} = \left(\frac{d\theta}{d\phi}\right)^2 I_\Theta(\theta).$$

3.5.1 Examples

Example 3.2 *(Normal mean.) Suppose $X_1, \ldots, X_n$ are independent variables distributed as* Normal (θ, σ^2)*, (σ^2 known).*

Then,

$$f(x|\theta) \propto \exp\{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\}$$

So,

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

$$\frac{d\ell(\theta)}{d\theta} = -\frac{1}{2\sigma^2}2[-\sum_{i=1}^n (x_i-\theta)] = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\theta),$$

and

$$\frac{d^2\ell(\theta)}{d\theta^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (-1) = -\frac{n}{\sigma^2}.$$

Then,

$$\begin{aligned} I(\theta) &= -E\left\{\frac{d^2\ell(\theta)}{d\theta^2}\right\} \\ &= E\left\{\frac{n}{\sigma^2}\right\} = \frac{n}{\sigma^2} \end{aligned}$$

Hence, $J(\theta) \propto 1$. Note that we have worked with the full likelihood here. However, we could have worked with the likelihood from a *single* observation x, and used the property that because of independence $I_n(\theta) = nI_1(\theta)$, where I_1 and I_n are the information from 1 and n independent values of x respectively. Thus we would (as required) obtain the same Jeffreys prior regardless of how many observations we subsequently make.

Example 3.3 *(Binomial sample.) Suppose* $X|\theta \sim \text{Binomial}\,(n, \theta)$.

Then,

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

and

$$\ell(\theta) = \log(f(x|\theta)) = x\log(\theta) + (n-x)\log(1-\theta) + c.$$

So,

$$\frac{d\ell(\theta)}{d\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

and

$$\frac{d^2\ell(\theta)}{d\theta^2} = \frac{-x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2},$$

and since $E(x) = n\theta$,

$$\begin{aligned} I(\theta) &= \frac{n\theta}{\theta^2} + \frac{(n-n\theta)}{(1-\theta)^2} = n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) \\ &= n\left(\frac{1-\theta+\theta}{\theta(1-\theta)}\right) = n\theta^{-1}(1-\theta)^{-1}, \end{aligned}$$

leading to

$$J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

which in this case is the proper distribution Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$.

3.6 Prior elicitation

There are no easy answers as to how best to elicit prior information, but the point of mentioning it here is that it is an important issue. It must be remembered however that nobody can make an infinite number of (or even many) accurate probability judgements. Thus, the specification of a prior distribution should be seen as an attempt to reconcile the beliefs an individual has into a unifying form. As we have seen, one approach is to take a particular family of distributions (the conjugate family, say), request prior information about some summary features such as prior mean and variance, and then choose as a prior the conjugate family member with those particular values. Generally though, it can be enormously difficult to reconcile an expert's (or even more so, experts') prior beliefs into a prior distribution.

3.7 Exercises

Exercise 3.1 *Which of the following densities belong to the exponential family (explain your working)?*

$$f_1(x|\theta) = \theta 2^\theta x^{-(\theta+1)} \text{ for } x > 2,$$

$$f_2(x|\theta) = \theta x^{\theta-1} \exp\{-x^\theta\} \text{ for } x > 0.$$

In each case calculate the conjugate prior if the density belongs to the exponential family.

Exercise 3.2 *Find the Jeffreys prior for θ in the geometric model:*

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

(Note $E(X) = 1/\theta$.)

Exercise 3.3 *Suppose x has the Pareto distribution* $\text{Pareto}\,(a, b)$*, where a is known but b is unknown. So,*

$$f(x|b) = ba^b x^{-b-1}; \qquad (x > a, \quad b > 0).$$

Find the Jeffreys prior and the corresponding posterior distribution for b.

Exercise 3.4 *(a) An observation x is made from the pdf $f(x|\theta)$. Define the Jeffreys' prior distribution $J_\theta(\theta)$ for the parameter θ in this context. State the invariance propert of this prior distribution, with respect to a parameter transformation $\phi = \phi(\theta)$.*

(b) Let $f(x|\theta) = \theta \exp(-x\theta)$ for $x, \theta \geq 0$. Derive Jeffreys' prior $J_\theta(\theta)$ for this model.

(c) Hence derive Jeffreys' prior $J_\phi(\phi)$ for this model when $\theta = \exp(\phi)$.

Exercise 3.5 *You are interested in estimating θ, the probability that a drawing pin will land point up. Your prior belief can be described by a mixture of Beta distributions:*

$$f(\theta) = \frac{\Gamma(a+b)}{2\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} + \frac{\Gamma(p+q)}{2\Gamma(p)\Gamma(q)}\theta^{p-1}(1-\theta)^{q-1}.$$

You throw a drawing pin n independent times, and observe x occasions on which the pin lands point up. Calculate the posterior distribution for θ.

Chapter 4

Multi–parameter problems

All the examples we have looked at so far involve just a single parameter — typically, a mean or variance of a population. Most statistical problems involve a statistical model which contains more than one unknown parameter. It might be the case that there is just one parameter of particular interest, but usually there will also be other parameters whose values are unknown.

The method of analysing multiparameter problems in Bayesian statistics is much more straightforward (at least in principle) than in the corresponding field of classical statistics. Indeed, there is absolutely no new theory required beyond what we have already looked at. We now have a vector $\theta = (\theta_1, \ldots, \theta_d)$ of parameters which we wish to make inferences about. We specify a prior (multivariate) distribution $f(\theta)$ for θ, and combine with a likelihood $f(x|\theta)$ via Bayes' theorem to obtain

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)d\theta}$$

as before. Of course, the posterior distribution will now also be a *multivariate* distribution. However, the simplicity of the Bayesian approach now means that inference about any subset of parameters within θ is obtained by straightforward probability calculations on this joint distribution.

The *conditional posterior distribution* of a component of θ, θ_i say, given the values of the remaining components θ_{-i} is given by

$$f_i(\theta_i \mid x, \theta_{-i}) \propto f(\theta \mid x),$$

where the values of θ_{-i} are held fixed. That is the conditional posterior distribution of θ_i is given by the joint posterior distribution of θ, $f(\theta \mid \mathbf{x})$, regarded as a function of θ_i alone with the other components θ_{-i} of θ fixed, normalised to be a density function as appropriate.

However, *exact Bayesian inference* about the scalar parameter θ_i can only be made from the posterior distribution integrated over θ_{-i},

$$f(\theta_i \mid \mathbf{x}) = \int f(\theta \mid \mathbf{x}) d\theta_{-i}.$$

This resulting *marginal posterior* of a given parameter of interest θ_i, after eliminating the nuisance parameters θ_{-i} by integration, can be used for drawing inferences about that parameter.

If marginalization is not possible, another approach which can be used to eliminate the nuisance parameters is to compute the posterior distribution of the parameter of interest conditioning on the maximum likelihood estimates of the other components of the parameter vector. This technique, which is not fully Bayesian, is called the *empirical Bayes* method, to be distinguished from fully Bayesian inferential methods.

Although no new theory is needed in multi-parametr problems, the increase in dimensionality does give rise to a number of practical problems:

1. **Prior specification.** Priors are now multivariate distributions. This means that the prior specification needs to reflect prior belief not just about each parameter individually, but also about dependence between different combinations of parameters (if one parameter is thought to be large, is it likely that another parameter should be correspondingly low?). Choosing suitable families of prior distributions and summarizing experts' prior information in this way is substantially more complicated.

2. **Computation.** Even in one–dimensional problems we saw the benefit of using conjugate families to simplify posterior analyses arising from the use of Bayes' theorem. With multivariate problems the integrals become even more difficult to evaluate. This makes the use of conjugate prior families even more valuable, and creates the need for numerical techniques to obtain inferences when conjugate families are either unavailable or inappropriate.

3. **Interpretation.** The entire posterior inference is contained in the posterior distribution, which will have as many dimensions as the variable θ. The structure of the posterior distribution may be highly complex, and it may require considerable skill (and a computer with good graphics facilities) to identify the most important relationships it contains.

Despite these practical aspects it is important to re–emphasise that precisely the same theory is being used for multi–parameter problems as one–dimensional problems. The Bayesian framework means that all inferences follow from elementary rules of probability.

4.1 Examples

Example 4.1 *Suppose a machine is either satisfactory ($x = 1$) or unsatisfactory ($x = 2$). The probability of the machine being satisfactory depends on the room temperature ($\theta_1 = 0$: cool, $\theta_1 = 1$: hot) and humidity ($\theta_2 = 0$: dry, $\theta_2 = 1$: humid). The probabilities of $x = 1$ are given in table 4.1. Furthermore, the joint prior distribution of (θ_1, θ_2) is given in table 4.2.*

$\Pr(x = 1\|\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.6	0.8
$\theta_2 = 1$	0.7	0.6

Table 4.1: Conditional probabilities of machine being satisfactory.

$\Pr(\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.3	0.2
$\theta_2 = 1$	0.2	0.3

Table 4.2: Prior probabilities of room conditions.

The joint posterior distribution can be calculated as in table 4.3. Hence, by summing across

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(x = 1\|\theta_1, \theta_2) \times \Pr(\theta_1, \theta_2)$	$\theta_2 = 0$	0.18	0.16
$= \Pr(x = 1, \theta_1, \theta_2)$	$\theta_2 = 1$	0.14	0.18
$\Pr(x = 1)$		0.66	
$\Pr(\theta_1, \theta_2\|x = 1)$	$\theta_2 = 0$	18/66	16/66
	$\theta_2 = 1$	14/66	18/66

Table 4.3: Posterior probabilities of room conditions.

margins we obtain the marginal posterior distributions:

$$\Pr(\theta_1 = 0) = 32/66, \quad \Pr(\theta_1 = 1) = 34/66$$

and

$$\Pr(\theta_2 = 0) = 34/66, \quad \Pr(\theta_2 = 1) = 32/66.$$

Example 4.2 *Suppose $Y_1 \sim \text{Poisson}(\alpha\beta)$ and $Y_2 \sim \text{Poisson}((1-\alpha)\beta)$ with Y_1 and Y_2 independent given α and β. Now, suppose our prior information for α and β can be expressed as: $\alpha \sim \text{Beta}(p, q)$ and $\beta \sim \text{Gamma}(p + q, 1)$ with α and β independent, for specified hyperparameters p and q.*

Note that it is necessary to specify the joint prior distribution of α and β. This is made more simple if we can assume prior independence of these parameters, because we need only then specify their marginal distributions. But this should only be done if it is a realistic assumption. (For example, we may have: β = expected number of motor vehicle accidents and α = expected proportion of those which are alcohol related).

Then we have the following likelihood:

$$f(y_1, y_2|\alpha, \beta) = \frac{\exp(-\alpha\beta)(\alpha\beta)^{y_1}}{y_1!} \times \frac{\exp(-(1-\alpha)\beta)[(1-\alpha)\beta]^{y_2}}{y_2!}$$

and the prior

$$f(\alpha, \beta) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\alpha^{p-1}(1-\alpha)^{q-1} \times \frac{e^{-\beta}\beta^{p+q-1}}{\Gamma(p+q)}.$$

Hence, by Bayes' theorem:

$$\begin{aligned} f(\alpha, \beta|y_1, y_2) &\propto e^{-\beta}\beta^{y_1+y_2}\alpha^{y_1}(1-\alpha)^{y_2}\alpha^{p-1}(1-\alpha)^{q-1}e^{-\beta}\beta^{p+q-1} \\ &= \beta^{y_1+y_2+p+q-1}e^{-2\beta}\alpha^{y_1+p-1}(1-\alpha)^{y_2+q-1} \end{aligned}$$

over the region $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq \infty$. This is the (joint) posterior distribution for α and β and contains all the information from the prior and data. In this particular case, the posterior factorises into functions of α and β. Therefore, we can write:

$$f(\alpha, \beta \mid y_1, y_2) \propto g(\alpha)h(\beta),$$

where

$$g(\alpha) = \alpha^{y_1+p-1}(1-\alpha)^{y_2+q-1}$$

and

$$h(\beta) = \beta^{y_1+y_2+p+q-1}e^{-2\beta}.$$

It follows, therefore, that the marginal posterior distributions are given by

$$f(\alpha|y_1, y_2) = \int_0^\infty f(\alpha, \beta|y_1, y_2)d\beta \;\propto\; g(\alpha),$$

and

$$f(\beta|y_1, y_2) = \int_0^1 f(\alpha, \beta|y_1, y_2)d\alpha \;\propto\; h(\beta).$$

That is, $\alpha|y_1, y_2 \sim \text{Beta}\,(y_1+p, y_2+q)$ and $\beta|y_1, y_2 \sim \text{Gamma}\,(y_1+y_2+p+q, 2)$.

4.2 Exercises

Exercise 4.1 *The quality of an electrical component is either excellent ($x = 1$), good ($x = 2$) or poor ($x = 3$). The probability of the various levels of quality depend on the factory of production*

($\theta_1 = 0$: factory A, $\theta_1 = 1$: factory B) and machine type ($\theta_2 = 0$: machine I, $\theta_2 = 1$: machine II, $\theta_2 = 3$: machine III). The probabilities of $x = 3$ are given in table 4.4. Furthermore, the

$\Pr(x = 3\|\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.2	0.3
$\theta_2 = 1$	0.4	0.1
$\theta_2 = 2$	0.5	0.2

Table 4.4: Conditional probabilities of $x = 3$.

joint prior distribution of (θ_1, θ_2) is given in table 4.5. Find the joint posterior distribution of

$\Pr(\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.1	0.2
$\theta_2 = 1$	0.2	0.3
$\theta_2 = 2$	0.1	0.1

Table 4.5: Prior probabilities of machine/factory combinations.

$\theta_1, \theta_2 | x = 3$ and each of the marginal distributions. Has having observed $x = 3$ affected which factory/machine combination is the most likely to have produced the component?

Exercise 4.2 *The random variable X is either 0 or 1 with probabilities depending on parameters α and β:*

$$\Pr(X = 0|\alpha, \beta) = \begin{cases} 0.2 & \alpha = \alpha_1, \beta = \beta_1 \\ 0.8 & \alpha = \alpha_1, \beta = \beta_2 \\ 0.9 & \alpha = \alpha_2, \beta = \beta_1 \\ 0.3 & \alpha = \alpha_2, \beta = \beta_2 \end{cases}$$

The joint prior is

$$\Pr(\alpha, \beta) = \begin{cases} 0.1 & \alpha = \alpha_1, \beta = \beta_1 \\ 0.5 & \alpha = \alpha_1, \beta = \beta_2 \\ 0.3 & \alpha = \alpha_2, \beta = \beta_1 \\ 0.1 & \alpha = \alpha_2, \beta = \beta_2 \end{cases}$$

Suppose we take 2 independent values of X, each of which happen to be 0.

(a) Write down the table of likelihoods.

(b) Obtain the joint posterior distribution.

(c) Obtain the marginal posterior distribution for α .

Exercise 4.3 *(a) Observations x_1 and x_2 are obtained of random variables X_1 and X_2, having Poisson distributions with respective means θ and $\phi\theta$, where ϕ is a known positive*

coefficient. Show, by evaluating the posterior density of θ, that the $\text{Gamma}\,(p, q)$ *family, of prior distributions of θ, is conjugate for this data model.*

What is the posterior mean for θ in the case that $p = 1$ and $q = 1$?

(b) Now suppose that ϕ is also an unknown parameter with prior density $f(\phi) = 1/(1+\phi)^2$, and independent of θ. Obtain the joint posterior distribution of θ and ϕ, and show that the marginal posterior distribution of ϕ is proportional to

$$\frac{\phi^{x_2}}{(1+\phi)^2(1+\phi+q)^{x_1+x_2+p}}.$$

Exercise 4.4 *Let $x_1, x_2, \ldots, x_n$ be a set of independent observations from the Normal distribution with mean θ and variance ϕ^{-1}.*

Show that the likelihood of the observations $x_1, x_2, \ldots, x_n$ is

$$L(\theta, \phi) \propto \phi^{n/2} \exp\left\{-\frac{\phi}{2}\left[s^2 + n(\bar{x}-\theta)^2\right]\right\},$$

where

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

Assume a Gamma prior for ϕ with parameters p and q and a Normal prior for θ with mean 0 and variance ϕ^{-1}. Write down the joint posterior for θ and ϕ (up to a normalising constant); what is the marginal posterior for ϕ and the conditional posterior for θ given ϕ?

Chapter 5

Summarizing posterior information

5.1 Decision theory

This is an extremely important area in its own right. Within this course we will only touch on the main issues.

Many problems in the real world are those of making decisions in the face of uncertainty: 'which political party will be best to vote for?'; 'should I accept one job offer or wait in the hope that I get offered a better job?'. All of statistical inference can also be thought of as decision making: having observed a particular set of data, what value should we decide to estimate a parameter by? There are many approaches to decision theory, but by far the most coherent is an approach based on Bayesian analysis. Indeed, it can be shown that if certain axioms are obeyed (that is, a number of common–sense decision rules are adopted) then a Bayesian analysis is the *only* logical approach to decision making. This is often used as an argument to justify Bayesian inference in preference to classical inference.

The elements needed to construct a decision problem are as follows:

1. A *parameter space* Θ which contains the possible *states of nature*;
2. A set A of *actions* which are available to the decision maker;
3. A *loss function* L, where $L(\theta, a)$ is the loss incurred by adopting action a when the true state of nature is θ.

We'll look at these terms in the context of a specific example:

Example 5.1 *First part. A public health officer is seeking a rational policy of vaccination against a relatively mild ailment which causes absence from work. Surveys suggest that 60% of the population are already immune. It is estimated that the money–equivalent of man–hours lost from failing to vaccinate a vulnerable individual is 20, that the unnecessary cost of vaccinating an immune person is 8, and that there is no cost incurred in vaccinating a vulnerable person or failing to vaccinate an immune person.*

So, for this particular case we have:

1. The parameter space $\Theta = \{\theta_1, \theta_2\}$, where θ_1 and θ_2 correspond to the individual being immune and vulnerable respectively;
2. The set of actions $A = \{a_1, a_2\}$ where a_1 and a_2 correspond to vaccinating and not vaccinating respectively;
3. The loss function $L(\theta, a)$ defined in Table 5.1

$L(\theta, a)$	θ_1	θ_2
a_1	8	0
a_2	0	20

Table 5.1: Loss function.

The decision strategy is then to evaluate the expected loss for each action and choose the action which has the minimum expected loss. This will be the same for each individual, in this case. Table 5.2 shows the calculation of the expected loss for each action, based upon the prior distribution of θ. The conclusion is that it does pay to vaccinate everyone. The cost (or loss) is 4.8 per individual.

$f(\theta)$	0.6	0.4	
$L(\theta, a)$	θ_1	θ_2	$\mathrm{E}\left[L(\theta, a)\right]$
a_1	8	0	$0.6 \times 8 + 0.4 \times 0 \;=\; 4.8$
a_2	0	20	$0.6 \times 0 + 0.4 \times 20 \;=\; 8.0$

Table 5.2: Expected loss function.

The assumption of zero loss in the case of 'correct' action is not a restrictive one. We can add or subtract one quantity to the first column in the loss function and another quantity to the second column without affecting the best action. So if the loss function had no zeros, we could adjust it so that the lowest value in each column was zero. All that matters for a given 'state of nature', θ, is the relative costs of the actions.

Suppose now that we had further information or data x available to us which reflected the value of θ. To be precise suppose that we have observed x from $f(x|\theta)$. Then we can replace $f(\theta)$ by the posterior $f(\theta|x)$ in the calculation of the expected loss. The best action will then depend on the particular outcome x.

Example 5.2 *Continuation. A simple skin test has been developed which, though not completely reliable, tends to indicate the immune status of the individual with respect to the disease. The probabilities of reaction are summarized in table 5.3*

			Immune Status	
			Immune θ_1	Vulnerable θ_2
Reaction	Negligible	x_1	0.35	0.09
	Mild	x_2	0.30	0.17
	Moderate	x_3	0.21	0.25
	Strong	x_4	0.14	0.49

Table 5.3: Probabilities of reaction given immune status.

Our general procedure is therefore to use Bayes' theorem to give us the posterior distribution $f(\theta|x)$. Then, for any particular action a, the *posterior expected loss* is

$$\rho(a, x) = \mathrm{E}\left[L(\theta, a)|x\right] = \int L(\theta, a) f(\theta|x) d\theta.$$

Having observed a particular value of x, we choose the action a which results in the lowest value of ρ. Writing $a = d(x)$, we call $d(x)$ the *Bayes decision rule.*

For our example, Table 5.4 shows the way we calculate the decision rule. We consider all the possible outcomes x, calculating for each of these the corresponding posterior $f(\theta|x)$. For each of these we next work out the expected posterior loss for each action. Finally we select the best action, that with the minimum expected posterior loss, for that outcome.

Table 5.5 shows the decision rule we obtain for this example.

So in summary, if either a negligible or mild reaction is observed, the Bayes decision is not to vaccinate, whereas if a moderate or strong reaction is observed, the decision is to vaccinate.

We can go one stage further and calculate the *risk* associated with this policy, by averaging across the uncertainty in the observations x. That is, we define the Bayes risk by:

$$BR(d) = \int \rho(d(x), x) f(x) dx$$

		Immune Status				
		Immune θ_1	Vulnerable θ_2			
Likelihoods	$f(x_1\|\theta)$	0.35	0.09			
	$f(x_2\|\theta)$	0.30	0.17			
	$f(x_3\|\theta)$	0.21	0.25			
	$f(x_4\|\theta)$	0.14	0.49			
Prior	$f(\theta)$	0.6	0.4			
Joints	$f(x_1, \theta)$	0.210	0.036	0.246	$f(x_1)$	
	$f(x_2, \theta)$	0.180	0.068	0.248	$f(x_2)$	
	$f(x_3, \theta)$	0.126	0.100	0.226	$f(x_3)$	
	$f(x_4, \theta)$	0.084	0.196	0.280	$f(x_4)$	
				Actions		
				a_1	a_2	
Posteriors	$f(\theta\|x_1)$	0.854	0.146	6.829	2.927	
	$f(\theta\|x_2)$	0.726	0.274	5.806	5.484	Expected
	$f(\theta\|x_3)$	0.558	0.442	4.460	8.847	Losses
	$f(\theta\|x_4)$	0.300	0.700	2.400	14.000	

Table 5.4: Tabulation of Bayesian decision analysis.

x	$d(x)$	$\rho(d(x), x)$
x_1	a_2	2.927
x_2	a_2	5.484
x_3	a_1	4.460
x_4	a_1	2.400

Table 5.5: Tabulation of Bayesian decision rule.

For our example this becomes the sum

$$BR(d) = \sum \rho(d(x), x) f(x) = 2.927 \times 0.246 + 5.484 \times 0.248 + 4.460 \times 0.226 + 2.400 \times 0.280 = 3.76$$

This is smaller than the least cost per individual, of 4.8, which we obtain by using the prior information alone, without the knowledge of x. This must always be the case. It does not mean that making the measurement is always worth while, because there is typically a fixed cost per individual associated with that. In this case there is a net benefit of measuring x if it costs less than the saving, of 4.8-3.76=1.04.

5.2 Point estimation

We've stressed throughout that the posterior distribution is a complete summary of the inference about a parameter θ. In essence, the posterior distribution *is* the inference. However, for some applications it is desirable (or necessary) to summarize this information in some way. In particular, we may wish to give a single 'best' estimate of the unknown parameter. (Note the distinction with classical statistics in which point estimates of parameters are the natural consequence of an inference, and it is reflecting uncertainty in that estimate which is more troublesome).

So, in the Bayesian framework, how do we reduce the information in a posterior distribution to give a single 'best' estimate? In fact, the answer depends on what we mean by 'best', and this in turn is specified by turning the problem into a decision problem. That is, we specify a loss function $L(\theta, a)$ which measures our perceived penalty in estimating θ by a. There are a range of natural loss functions we could use, and the particular choice for any specified problem will depend on the context. The most commonly used are:

1. *Squared Error (or Quadratic) loss*: $L(\theta, a) = (\theta - a)^2$;
2. *Absolute Error loss*: $L(\theta, a) = |\theta - a|$;
3. *0—1 loss*:

$$L(\theta, a) = \begin{cases} 0 & \text{if } |\theta - a| \leq \epsilon \\ 1 & \text{if } |\theta - a| > \epsilon \end{cases}$$

In each of these cases, by minimizing the posterior expected loss, we obtain simple forms for the Bayes decision rule, which is taken to be the point estimate of θ for that particular choice of loss function.

5.2.1 Squared error loss

In this case we can simplify $\rho(a, x) = \mathrm{E}\left[(\theta - a)^2 | x\right]$ by letting $\mu = \mathrm{E}\left(\theta | x\right)$ and expanding:

$$\begin{aligned} \mathrm{E}\left[(\theta - a)^2 | x\right] &= \mathrm{E}\left\{\left[(\theta - \mu) + (\mu - a)\right]^2 | x\right\} \\ &= \mathrm{E}\left[(\theta - \mu)^2 | x\right] + (\mu - a)^2 + 2\,\mathrm{E}\left[(\theta - \mu) | x\right](\mu - a) \\ &= \mathrm{Var}[\theta | x] + (\mu - a)^2 \end{aligned}$$

On the right, the first term no longer depends on a, and the second term attains its minimum of zero by taking $a = \mu$. In summary, the posterior expected squared error loss has its minimum value of $\text{Var}[\theta|x]$, the posterior variance of θ, when $a = \text{E}(\theta|x)$, the posterior expectation of θ.

5.2.2 Absolute error loss

We show that in this case the minimum posterior expected loss is obtained by taking $a = m$, the median of the posterior distribution $f(\theta|x)$. We assume that this is unique, and is defined by

$$\Pr(\theta < m|x) = \Pr(\theta > m|x) = \tfrac{1}{2}.$$

To prove the result note first that the function

$$s(\theta) = \begin{cases} -1, \text{ for } \theta < m \\ +1, \text{ for } \theta > m \end{cases}$$

has the property

$$\begin{aligned} \text{E}[s(\theta) \mid x] &= -\int_{-\infty}^{m} f(\theta \mid x)d\theta + \int_{m}^{\infty} f(\theta \mid x)d\theta \\ &= -\Pr(\theta < m \mid x) + \Pr(\theta > m \mid x) = 0. \end{aligned}$$

Now consider $L(\theta, a) - L(\theta, m) = |\theta - a| - |\theta - m|$ for some $a < m$.

If $\theta < a$: $L(\theta, a) - L(\theta, m) = -\theta + a + \theta - m = a - m = (m - a)s(\theta)$

If $\theta > m$: $L(\theta, a) - L(\theta, m) = -a + \theta - \theta + m = -a + m = (m - a)s(\theta)$

If $a < \theta < m$: $L(\theta, a) - L(\theta, m) = -a + \theta + \theta - m = 2\theta - a - m > (m - a)s(\theta)$

It can be seen that $L(\theta, a) - L(\theta, m)$ is greater than $(m - a)s(\theta)$ so

$$\text{E}[L(\theta, a) - L(\theta, m)|x] > (m - a)\text{E}[s(\theta)|x] = 0.$$

So $\text{E}[L(\theta, a)|x] > \text{E}[L(\theta, m)|x]$. This also holds by a similar argument when $a > m$, so $\text{E}[L(\theta, a)|x]$ is a minimum when $a = m$.

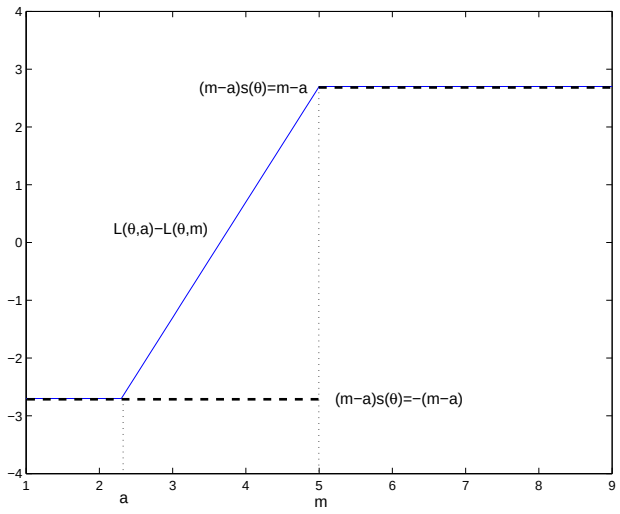

Figure 5.1: Plot of $L(\theta, a) - L(\theta, m)$ and $(m - a)s(\theta)$.

5.2.3 0—1 loss

Clearly in this case

$$\rho(a, x) = \Pr\{|\theta - a| > \epsilon | x) = 1 - \Pr\{|\theta - a| \leq \epsilon | x).$$

Consequently, if we define a *modal interval of length* 2ϵ as the interval $[\theta - \epsilon, \theta + \epsilon]$ which has highest probability, then the Bayes estimate is the midpoint of the interval with highest probability. By choosing ϵ arbitrarily small, this procedure will lead to the posterior mode as the Bayes estimate with this particular loss function.

Example 5.3 *If the posterior density for* θ *is*

$$f(\theta|x) = 1 \text{ for } 0 \leq \theta \leq 1,$$

calculate the best estimator of $\phi = \theta^2$ *with respect to quadratic loss.*

The best estimator of ϕ with respect to quadratic loss is

$$\mathrm{E}\left(\phi \mid x\right) = \mathrm{E}\left(\theta^2 \mid x\right) = \int_0^1 \theta^2 d\theta = \left[\frac{\theta^3}{3}\right]_0^1 = \frac{1}{3}.$$

5.2.4 Conclusion

In conclusion then, the important point is that in the Bayesian framework a point estimate of a parameter is a single summary statistic of the posterior distribution. By defining the quality of

an estimator through a loss function, the decision theory methodology leads to optimal choices of point estimates. In particular, the most natural choices of loss function lead respectively to the posterior mean, median and mode as optimal point estimators.

5.3 Credibility intervals - Credibility Regions

The idea of a credibility interval is to give an analogue of a confidence interval in classical statistics. The reasoning is that point estimates give no measure of accuracy, so it is preferable to give an interval within which it is 'likely' that the parameter lies. This causes problems in classical statistics since parameters are not regarded as random, so it is not possible to give an interval with the interpretation that there is a certain probability that the parameter lies in the interval. (Instead, confidence intervals have the interpretation that if the sampling were repeated, there is a specified probability that the interval so obtained would contain the parameter — it is the interval which is random and not the parameter.)

There is no such difficulty in the Bayesian approach because parameters are treated as random. Thus, a region $C_\alpha(x)$ is a $100(1-\alpha)\%$ *credible region* for θ if

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

That is, there is a probability of $1-\alpha$, based on the posterior distribution, that θ lies in $C_\alpha(x)$.

Some Bayesians argue that credible intervals have little value since it is the entire posterior distribution which contains the information for inference and that credible intervals have only been proposed in order to give something comparable to confidence intervals.

One difficulty with credible intervals (in common with confidence intervals) is that they are not uniquely defined. Any region with probability $1-\alpha$ will do. Since we want the interval to contain only the 'most plausible' values of the parameter, it is usual to impose an additional constraint which is that the width of the interval should be as small as possible. This amounts to an interval (or region) of the form

$$C_\alpha(x) = \{\theta : \ f(\theta|x) \geq \gamma\}$$

where γ is chosen to ensure that

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

Such regions are called *highest posterior density regions*. If the posterior density is unimodal, the highest posterior density regions are intervals of the form (a, b) (for example, see Figure 5.2).

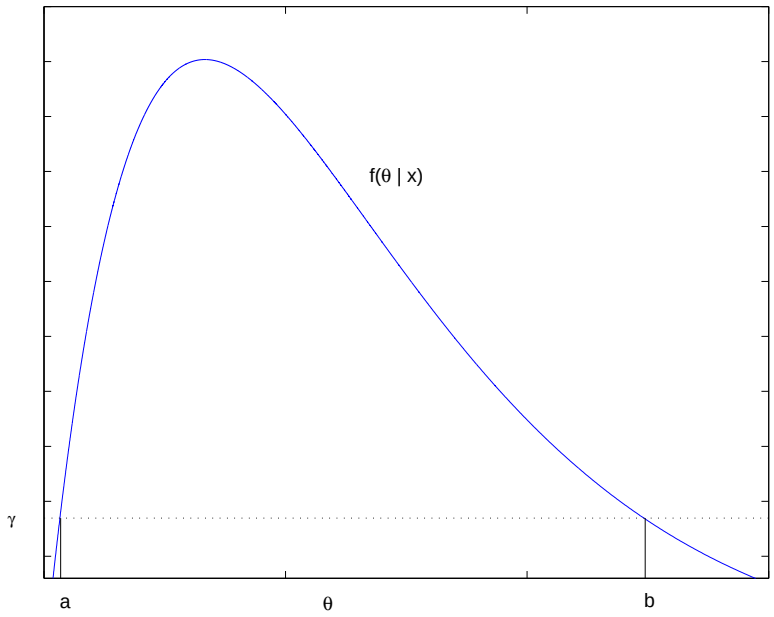

Figure 5.2: Highest posterior density region of a unimodal posterior distribution. The region is an interval of the form (a, b).

In general, these intervals have to be found numerically, though for most standard univariate posterior distributions values are tabulated for a range of values of α. (In passing, note that there is the usual trade–off in making the choice an appropriate α: small values of α will give large intervals; large values give intervals which the parameter has only a low probability of lying within.)

Common Mistake 4 *You should ALWAYS sketch the posterior density before trying to calculate the HPD region.*

Example 5.4 *(Normal mean). Let $X_1, \ldots, X_n$ be independent variables from* Normal (θ, σ^2) *(σ^2 known) with a prior for θ of the form $\theta \sim$ Normal (b, d^2).*

With this construction we obtained the posterior:

$$\theta|x \sim \text{Normal}\,(\mu, s^2)$$

where $\mu = \frac{\frac{b}{d^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$ and $s^2 = \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$.

Now, since the normal distribution is uni–modal and symmetric the highest posterior density regions are symmetric intervals of the form $(\mu - c, \mu + c)$ (see Figure 5.3). It follows that the

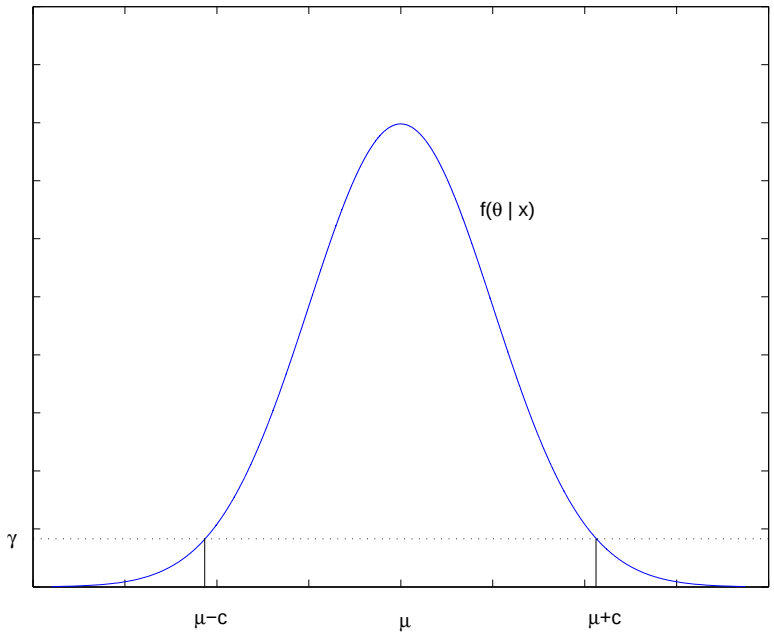

Figure 5.3: Highest posterior density region of a Normal (μ, s^2) posterior distribution. The region is an interval of the form $(\mu - c, \mu + c)$.

$100(1-\alpha)\%$–highest posterior density region for θ is:

$$\left(\frac{\frac{b}{d^2} + \frac{n\overline{x}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}\right) \pm z_{\alpha/2} \left(\frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}\right)^{\frac{1}{2}},$$

where $z_{\alpha/2}$ is the appropriate percentage point of the standard normal, Normal $(0, 1)$ distribution.

Notice, moreover, that as $n \to \infty$ this interval becomes

$$\overline{x} \pm z_{\alpha/2}\sigma/\sqrt{n},$$

which is precisely the $100(1-\alpha)\%$ confidence interval for θ obtained in classical inference. In this special case then, the Bayesian credible interval and the confidence interval are identical, though their *interpretations* are quite different.

Example 5.5 *Let $X_1, \ldots, X_n$ be independent variables from* Normal (θ, ϕ) *(ϕ unknown) and assume a reference prior of the form*

$$f(\theta, \phi) \propto \frac{1}{\phi}; \qquad -\infty < \theta < \infty,\ 0 < \phi.$$

This leads to the marginal posterior distributions:

$$t = \frac{\theta - \overline{x}}{s/\sqrt{n}} \sim t_{n-1}$$

and

$$S^2/\phi \sim \chi^2_{n-1}.$$

So, because of the symmetry of the t–distribution, the $100(1-\alpha)\%$ credible interval for θ is:

$$\overline{x} \pm t_{n-1,\alpha/2} s/\sqrt{n}$$

where $t_{n-1,\alpha/2}$ is the appropriate percentage point from the t_{n-1} distribution.

On the other hand, the credible interval for ϕ is slightly more troublesome. Since $S^2/\phi \sim \chi^2_{n-1}$, it follows that $\phi/S^2 \sim 1/\chi^2_{n-1}$, a so–called inverse chi–squared distribution. Critical points of highest posterior density intervals for a variety of values of α are available in tables.

Example 5.6 *Suppose* $x \sim \text{Binomial}\,(n,\theta)$ *with the prior*

$$\theta \sim \text{Beta}\,(p,q).$$

This gives the posterior distribution

$$\theta|x \sim \text{Beta}\,(p+x, q+n-x)$$

Thus, the $100(1-\alpha)\%$ highest posterior density interval $[a,b]$ satisfies:

$$\frac{1}{\text{B}(p+x,q+n-x)} \int_a^b \theta^{p+x-1}(1-\theta)^{q+n-x-1} d\theta = 1-\alpha,$$

and

$$\frac{1}{\text{B}(p+x,q+n-x)} a^{p+x-1}(1-a)^{q+n-x-1} = \frac{1}{\text{B}(p+x,q+n-x)} b^{p+x-1}(1-b)^{q+n-x-1} = \gamma.$$

Generally, this has to be solved numerically.

5.4 Hypothesis testing

Hypothesis tests are decisions of the form of choosing between two different hypotheses:

$$H_0 : \ \theta \in \Omega_0,$$
$$H_1 : \ \theta \in \Omega_1.$$

In the simplest case where Ω_1 and Ω_2 consist of single points, the test is of the form

$$H_0 : \ \theta = \theta_0,$$
$$H_1 : \ \theta = \theta_1.$$

The classical approach to this problem is usually to base the test on the *likelihood ratio*:

$$\lambda = \frac{f(x|\theta_1)}{f(x|\theta_0)}.$$

Large values of λ indicate that the observed data x is more likely to have occurred if θ_1 is the true value of θ than if θ_0 is. In the Bayesian view of things, we should also bring to bear the prior information we have about θ. Therefore, we may compute the posterior probabilities of θ_1 and θ_0:

$$\begin{aligned} f(\theta_1|x) &= \frac{f(\theta_1)f(x|\theta_1)}{f(\theta_0)f(x|\theta_0) + f(\theta_1)f(x|\theta_1)} \\ f(\theta_0|x) &= 1 - f(\theta_1|x). \end{aligned}$$

Hence it is natural to base test considerations on the relative posterior probabilities of the hypothesised values. That is we look at

$$\lambda_B = \frac{f(\theta_1|x)}{f(\theta_0|x)} = \frac{f(\theta_1)f(x|\theta_1)}{f(\theta_0)f(x|\theta_0)}. \tag{5.1}$$

This is usually called the *posterior odds*. Observe in particular that there is no requirement to calculate normalizing factors since the same factor would appear on the numerator and the denominator. Again, large values of λ_B would indicate in favour of H_1.

There is a related concept known as the *Bayes Factor*. We can see from equation 5.1 that the posterior odds is the product of the prior odds times the likelihood ratio. In this context, the likelihood ratio is termed the Bayes factor. That is

$$BF = \frac{f(x|\theta_1)}{f(x|\theta_0)} = \frac{f(\theta_1|x)/f(\theta_0|x)}{f(\theta_1)/f(\theta_0)}$$

The purpose of focusing on the Bayes factor is that it is a measure of the weight of information contained in the data in favour of H_1 over H_0. If the Bayes factor is sufficiently large, then it will overcome any prior preference we might have had for H_0 so that our posterior preference might be for H_1.

In the general case that we are interested in testing the hypotheses:

$$\begin{aligned} H_0 &: \ \theta \in \Omega_0, \\ H_1 &: \ \theta \in \Omega_1, \end{aligned}$$

we can still calculate the posterior probabilities of the two hypotheses, after specifying prior probabilities, $f(\theta \in \Omega_0)$ and $f(\theta \in \Omega_1)$, on the hypotheses. Then we have

$$f(\theta \in \Omega_1|x) = \frac{f(\theta \in \Omega_1)f(x|\theta \in \Omega_1)}{f(\theta \in \Omega_0)f(x|\theta \in \Omega_0) + f(\theta \in \Omega_1)f(x|\theta \in \Omega_1)},$$

where

$$f(x|\theta \in \Omega) = \int_{\Omega} f(\theta) f(x|\theta) d\theta.$$

Obviously, it is straightforward to generalise the above testing approach to the case of testing more than two hypotheses.

5.5 Bayesian Model Comparison

Comparing a number of competing models for a given set of observed data is an important area within Bayesian decision making. Before introducing the problem of model comparison, let us define the marginal likelihood of a given model.

The *marginal likelihood* or *evidence* $f(x)$ of a given model $f(x \mid \theta)$ is the marginal distribution of the data under that model. It is obtained by integrating the product of the likelihood times a prior distribution $f(\theta)$ on the model parameters θ over θ:

$$f(x) = \int f(x \mid \theta) f(\theta) d\theta.$$

That is $f(x)$ is the normalising constant of the posterior distribution of θ, given by

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

Note that all constants in the likelihood and the prior are included in the product.

Equivalently, the marginal likelihood is defined as the expectation of the likelihood with respect to the prior distribution $f(\theta)$. In order to be able to evaluate the integral(s) involved in the calculation of $f(x)$ we need to choose carefully a proper (preferably conjugate) prior on θ.

Marginal likelihoods play an important role in *Bayesian model comparison*. Consider a number of competing models $M_1, \ldots, M_k$, parameterised respectively by $\theta_1, \ldots, \theta_k$, for an observed data set. In the presence of uncertainty about the correct model, Bayesian inference involves:

1. Evaluation of the posterior probability $\Pr(M_j \mid x)$ of each model M_j, $j = 1, \ldots, k$.
2. Evaluation of the posterior distribution $f(\theta_j \mid x, M_j)$ of the parameters θ_j of model M_j, $j = 1, \ldots, k$.

After specifying prior model probabilities, $\Pr(M_j)$, for all competing models and carefully choosing proper prior distributions for the model specific parameters, $f(\theta_j \mid M_j)$, $j = 1, \ldots, k$, posterior inferences are obtained as follows.

1. The posterior probability of model M_j is calculated using Bayes therom as

$$\Pr(M_j \mid x) = \frac{\Pr(M_j) f(x \mid M_j)}{\sum_{i=1}^{k} \Pr(M_i) f(x \mid M_i)}, \quad j = 1, \ldots, k,$$

   where $f(x \mid M_j)$ is the marginal likelihood of model M_j.

2. The posterior distribution of the parameters θ_j of model M_j is given by Bayes therom as

$$f(\theta_j \mid x, M_j) = \frac{f(\theta_j \mid M_j) f(x \mid \theta_j, M_j)}{f(x \mid M_j)}, \quad j = 1, \ldots, k.$$

Example 5.7 *To demonstrate Bayesian model comparison, we'll look at a fairly intricate example concerning the change point in a Poisson process. The data consists of a series relating to the number of British coal mining disasters per year, over the period 1851 — 1962. A plot of these data is given in Figure 5.4. From this plot it does seem to be the case that there has been a reduction in the rate of disasters over the period.*

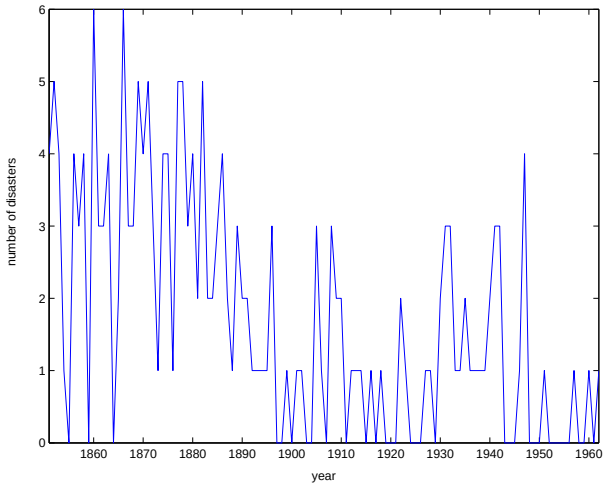

Figure 5.4: Time series of counts of coal mine disasters.

For the coal-mining disasters data we may, therefore, consider two models:

M1 *each x_i is an independent draw from a Poisson random variable with mean θ;*

M2 *for $i \leq t$, x_i is an independent draw from a Poisson random variable with mean θ_1, and for $i > t$, x_i is an independent draw from a Poisson random variable with mean θ_2.*

5.6 Exercises

Exercise 5.1 *An art 'expert' has undergone a test of her reliability in which she has separately pronounced judgement — 'genuine' or 'counterfeit' — on a large number of art subjects of know origin. From these it appears that she has probability 0.8 of detecting a counterfeit and probability 0.7 of recognising a genuine object. I have been offered an objet d'art at what seems a bargain price of £100. If it is not genuine, then it is worth nothing. If it is genuine I believe I can sell it immediately for £300. I believe there is a 0.5 chance that the object is genuine. The expert charges £30 for her services. Is it to my advantage to pay her for an assessment?*

Exercise 5.2 *Consider a decision problem with two actions, a_1 and a_2, and a loss function which depends on a parameter θ, with $0 \leq \theta \leq 1$. The loss function is*

$$L(\theta, a) = \begin{cases} 0 & a = a_1, \\ 2 - 3\theta & a = a_2. \end{cases}$$

Assume a Beta$(1,1)$ prior for θ, and an observation $X \sim$ Binomial(n, θ). The posterior distribution is Beta$(x+1, n-x+1)$.

Calculate the expected loss of each action and the Bayes rule. (You may use the fact that if $\theta \sim$ Beta(p,q), then $E(\theta) = p/(p+q)$.)

Exercise 5.3 *(a) For a parameter θ with a posterior distribution described by the* $\mathrm{Beta}\,(P,Q)$ *distribution, find the posterior mode in terms of P and Q and compare it with the posterior mean. (b) Repeat part (a) for the case when the posterior distribution of θ is* $\mathrm{Gamma}\,(P,Q)$.

Exercise 5.4 *The parameter θ has a* $\mathrm{Beta}\,(3,2)$ *posterior density. Show that the interval [5/21,20/21] is a 94.3% highest posterior density region for θ.*

Exercise 5.5 *A parameter θ has posterior expected value μ and variance v. Let $\phi = \theta^2$. Show that the estimate of ϕ which has minimum expected posterior squared error loss is given by $a = \mu^2 + v$.*

Exercise 5.6 *A parameter θ has posterior density that is* $\mathrm{Gamma}\,(1,1)$. *Calculate the 95% highest posterior density region for θ. Now consider the transformation of the parameter given by $\phi = \sqrt{2\theta}$. Obtain the posterior density of ϕ and explain why the highest posterior density region for ϕ is not obtained by transforming the interval for θ in the same way.*

Exercise 5.7 *In a decision theory setting, let $L(\theta, a)$ be the loss incurred by taking action a when the value of an unknown parameter is θ, and let $f(\theta|x)$ be the posterior distribution of θ given the measurement of a relevant variable x.*

(a) Determine the Bayes decision rule $d(x)$, and the corresponding losses, for the discrete example, where the loss function and posterior distribution are given in the tables below.

Losses	θ_1	θ_2	θ_3
a_1	*5*	*10*	*25*
a_2	*15*	*15*	*5*

	θ_1	θ_2	θ_3
$f(\theta\|x_1)$	*0.5*	*0.3*	*0.2*
$f(\theta\|x_2)$	*0.1*	*0.5*	*0.4*

(b) What further information would be required to determine the Bayes risk for this problem, and whether the cost of measuring x was worthwhile?

Exercise 5.8 *Consider a sample $x_1, \ldots, x_n$ consisting of independent draws from a Poission random variable with mean θ. Consider the hypothesis test, with Null hypothesis*

$$H_0 : \theta = 1$$

against an alternative hypothesis

$$H_1 : \theta \neq 1$$

Assume a prior probability of 0.95 *for H_0 and a Gamma prior*

$$f(\theta) = \frac{q^p}{\Gamma(p)} \theta^{p-1} \exp\{-q\theta\},$$

under H_1.

(a) Calculate the posterior probability of H_0.

(b) Assume $n = 10$, $\sum_{i=1}^{n} x_i = 20$, and $p = 2q$. What is the posterior probability of H_0 for each of $p = 2, 1, 0.5, 0.1$. What happens to this posterior probability as $p \to 0$?

Chapter 6

Prediction

6.1 The predictive distribution

So far we have focused on parameter estimation. That is, we have specified a probability model to describe the random process which has generated a set of data, and have shown how the Bayesian framework combines sample information and prior information to give parameter estimates in the form of a posterior distribution. Commonly the purpose of formulating a statistical model is to make *predictions* about future values of the process. This is handled much more elegantly in Bayesian statistics than in the corresponding classical theory. The essential point is that in making predictions about future values on the basis of an estimated model there are two sources of uncertainty:

- Uncertainty in the parameter values which have been estimated on the basis of past data; and
- Uncertainty due to the fact that any future value is itself a random event.

In classical statistics it is usual to fit a model to the past data, and then make predictions of future values on the assumption that this model is correct, the so–called *estimative* approach. That is, only the second source of uncertainty is included in the analysis, leading to estimates which are believed to be more precise than they really are. There is no completely satisfactory way around this problem in the classical framework since parameters are not thought of as being random.

Within Bayesian inference it is straightforward to allow for both sources of uncertainty by simply averaging over the uncertainty in the parameter estimates, the information of which is completely

contained in the posterior distribution.

So, suppose we have past observations $x = (x_1, \ldots, x_n)$ of a variable with density function (or likelihood) $f(x|\theta)$ and we wish to make inferences about the distribution of a future value of a random variable Y from this same model. With a prior distribution $f(\theta)$, Bayes' theorem leads to a posterior distribution $f(\theta|x)$. Then the *predictive density function* of y given x is:

$$f(y|x) = \int f(y|\theta)f(\theta|x)d\theta. \tag{6.1}$$

Thus the predictive density, evaluated at a particular value of y, is the integral of the likelihood of y times the posterior. The result can also be written as the expectation of the predictive density with respect to the posterior distribution of θ:

$$f(y|x) = \mathrm{E}\left[f(y|\theta)|x\right].$$

Again it is important to notice that this result is simply constructed from the usual laws of probability manipulation, and has a straightforward interpretation itself in terms of probabilities. The derivation must, however be presented carefully, with due emphasis on the assumptions. In the following, Y need not come from the same distribution as the observations x. It is important however that, *supposing θ to be known*, we assume that Y is independent of x. This enables us to write the joint density of Y and x, given θ as the product of their densities:

$$f(y, x|\theta) = f(y|\theta)f(x|\theta),$$

and from this we get the joint density of y, x and θ:

$$f(y, x, \theta) = f(y|\theta)f(x|\theta)f(\theta).$$

Then

$$f(y, \theta|x) = f(y|\theta)f(x|\theta)f(\theta)/f(x) = f(y|\theta)f(\theta|x),$$

and finally, integrating out θ we get the result

$$f(y|x) = \int f(y|\theta)f(\theta|x)d\theta.$$

The corresponding approach in classical statistics would be, for example, to obtain the maximum likelihood estimate $\hat{\theta}$ of θ and to base inference on the distribution $f(y|\hat{\theta})$, the *estimative* distribution. To emphasise again, this makes no allowance for the variability incurred as a result of estimating θ, and so gives a false sense of precision (the predictive density $f(y|x)$ is more variable by averaging across the posterior distribution for θ).

Common Mistake 5 *You CANNOT remove a constant of proportionality in $f(y|\theta)$.*

Common Mistake 6 *It is usually simplest to use the (normalised) posterior distribution in*

$$f(y|x) = \int f(y|\theta) f(\theta|x) d\theta.$$

(If you use the posterior density up to a constant of proportionality, then the answer will be for $f(y|x)$ up to a constant of proportionality.)

Though simple in principle, computation can be difficult. However, many of the standard conjugate families of prior–likelihood forms do lead to tractable forms for the predictive distribution.

Example 6.1 *(Binomial sample.) Suppose we have made an observation $x \sim \text{Binomial}\,(n, \theta)$ and our (conjugate) prior for θ is $\theta \sim \text{Beta}\,(p, q)$. Then, we have shown, the posterior for θ is given by:*

$$\theta|x \sim \text{Beta}\,(p + x, q + n - x).$$

Now, suppose we intend to make a further N observations in the future, and we let z be the number of successes in those N trials, so that $z|\theta \sim \text{Binomial}\,(N, \theta)$.

So, we have the likelihood for our future observation:

$$f(z|\theta) = \binom{N}{z} \theta^z (1-\theta)^{N-z}.$$

Before continuing with the general case let us look at the simple special case when $N = 1$, that is z is the outcome of a single Bernoulli trial. Writing, as before $P = p + x$, $Q = q + n - x$, the predictive distribution of z given x is

z	0	1
$f(z \mid \theta)$	$1 - \theta$	θ
$\text{E}\,(f(z \mid \theta) \mid x)$	$1 - \frac{P}{P+Q}$	$\frac{P}{P+Q}$

Therefore, the distribution of $z \mid x$ is Bernoulli with probability of success $\frac{P}{P+Q}$, the posterior mean of θ.

In the general case, for $z = 0, 1, \ldots, N$,

$$\begin{aligned} f(z|x) &= \int_0^1 \binom{N}{z} \theta^z (1-\theta)^{N-z} \times \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{\text{B}(p+x, q+n-x)} d\theta \\ &= \binom{N}{z} \frac{1}{\text{B}(P,Q)} \int_0^1 \theta^{P+z-1}(1-\theta)^{Q+N-z-1} \\ &= \binom{N}{z} \frac{\text{B}(P+z, Q+N-z)}{B(P,Q)}. \end{aligned}$$

This is, in fact, known as a Beta–binomial distribution.

Example 6.2 *(Gamma sample.) As in Chapter 2, suppose $X_1, \ldots X_n$ are independent variables having the* $\text{Gamma}\,(k, \theta)$ *distribution, where k is known, and we use the conjugate prior $\theta \sim$* $\text{Gamma}\,(p, q)$*:*

$$f(\theta) \propto \theta^{p-1} \exp\{-q\theta\}$$

leading via Bayes' theorem to $\theta|x \sim \text{Gamma}\,(p + nk, q + \Sigma x_i) = \text{Gamma}\,(G, H)$ say.

The likelihood for a future observation y is

$$f(y|\theta) = \frac{\theta^k y^{k-1} \exp\{-\theta y\}}{\Gamma(k)}$$

and so

$$\begin{aligned} f(y|x) &= \int_0^\infty \frac{\theta^k y^{k-1} \exp\{-\theta y\}}{\Gamma(k)} \times \frac{H^G \theta^{G-1} \exp\{-H\theta\}}{\Gamma(G)} d\theta \\ &= \frac{H^G y^{k-1}}{\Gamma(k)\Gamma(G)} \int_0^\infty \theta^{k+G-1} \exp\{-\theta(y+H)\} d\theta \\ &= \frac{H^G y^{k-1}}{\Gamma(k)\Gamma(G)} \frac{\Gamma(k+G)}{(y+H)^{k+G}} = \frac{H^G y^{k-1}}{\text{B}(k, G)(H+y)^{G+k}}, \quad y > 0, \end{aligned}$$

since $\frac{(y+H)^{k+G}}{\Gamma(k+G)} \theta^{k+G-1} \exp\{-\theta(y+H)\}$ is a pdf, and therefore

$$\int_0^\infty \frac{(y+H)^{k+G}}{\Gamma(k+G)} \theta^{k+G-1} \exp\{-\theta(y+H)\} d\theta = 1.$$

We can relate $f(y \mid x)$ to a standard distribution by writing

$$Y = (H\nu_1/\nu_2) F_{\nu 1, \nu 2},$$

where $\nu_1 = 2k$ and $\nu_2 = 2G$ and $F_{\nu 1, \nu 2}$ has the Fisher 'F' distribution.

Common Mistake 7 *Generally, to calculate these predictive distributions requires integrating a function of θ that is proportional to a standard density function. We can use the fact that we know the normalising constant of the density function to help us calculate the integral. When recognising the density function, remember that it will be density for θ (as opposed to θ being a known parameter in the density). (See common mistake 3).*

6.2 Exercises

Exercise 6.1 *A random sample $x_1, \ldots, x_n$ is observed from a* Poisson (θ) *distribution. The prior on θ is* Gamma (g, h) *(g and h are integers). Show that the predictive distribution for a future observation, y, from this* Poisson (θ) *distribution is*

$$f(y|x) = \binom{y+G-1}{G-1}\left(\frac{1}{1+H}\right)^y\left(1-\frac{1}{1+H}\right)^G; \qquad y = 0, 1, \ldots,$$

for some value of G and H. What is this distribution? (Use the fact that $\Gamma(n+1) = n!$ if n is an integer.)

Exercise 6.2 *The distribution of flaws along the length of an artificial fibre follows a Poisson process, so that the number of flaws in a length l of the fibre is* Poisson $(l\theta)$. *Very little is known about θ, so the Jeffrey's prior is used for it. The number of flaws obtained in 5 fibres of lengths 10, 15, 25, 30 and 40 metres respectively were 3, 2, 7, 6 and 10. Find the predictive distribution for the number of flaws in another fibre of length 60 metres.*

Exercise 6.3 *A random sample $x_1, \ldots, x_n$ is observed from a* Normal (θ, σ^2) *distribution with σ^2 known, and a normal prior for θ is assumed, leading to a posterior distribution* Normal (B, D^2) *for θ. Show that the predictive distribution for a further observation, y, from the* Normal (θ, σ^2) *distribution, is* Normal $(B, D^2 + \sigma^2)$.

Exercise 6.4 *The weights of items from a particular production process are independently and identically distributed, each with a $N(\theta, 4)$ distribution. The production manager believes that θ varies from batch to batch according to a $N(110, 0.4)$ distribution. A sample of 5 items is randomly selected, yielding the measurements:*

$$108.0, 109.0, 107.4, 109.6, 112.0.$$

Derive the posterior distribution for θ. Find also the predictive distribution for a) the weight of one further item from the batch; b) the sample mean of the weights of m further items from the batch.

Exercise 6.5 *Observations $x = (x_1, x_2, \ldots, x_n)$ are made of independent random variables $X = (X_1, X_2, \ldots, X_n)$ with X_i having uniform distribution*

$$f(x_i|\theta) = \frac{1}{\theta}; \quad 0 \leq x_i \leq \theta.$$

Assume that θ has an improper prior distribution

$$f(\theta) = \frac{1}{\theta}; \quad \theta \geq 0.$$

(a) Show that the posterior distribution of θ is given by

$$f(\theta|x) = \frac{nM^n}{\theta^{n+1}}; \quad \theta \geq M,$$

where $M = \max(x_1, x_2, \ldots, x_n)$.

(b) Show that θ has posterior expectation

$$\mathrm{E}\,(\theta|x) = \frac{n}{n-1}M.$$

(c) Verify the posterior probability:

$$\Pr(\theta > t\,M|x) = \frac{1}{t^n} \quad \textit{for any } t \geq 1.$$

(d) A further, independent, observation Y is made from the same distribution as X. Show that the predictive distribution of Y has density

$$f(y|x) = \frac{1}{M}\left(\frac{n}{n+1}\right)\frac{1}{[\max(1, y/M)]^{n+1}}; \quad y \geq 0.$$

(e) Sketch this density in the case $n = 1$ and $M = 1$.

Chapter 7

Asymptotics

7.1 Introduction

Looking back at the conjugate analysis for the Normal mean θ with $X_1 \ldots X_n \sim \text{Normal}\left(\theta, \tau^{-1}\right)$, we obtained for a prior $\theta \sim N(b, c^{-1})$

$$\theta|x \sim \text{Normal}\left(\tfrac{cb+n\tau\overline{x}}{c+n\tau}, \tfrac{1}{c+n\tau}\right)$$

Now, as $n \to \infty$, this becomes:

$$\theta|x \sim \text{Normal}\left(\overline{x}, 1/(n\tau)\right) = \text{Normal}\left(\overline{x}, \sigma^2/n\right)$$

Thus, as n becomes large, the effect of the prior disappears, and the posterior is determined solely by the data. Moreover, the posterior distribution becomes increasingly more concentrated around $\overline{x}$, which by the strong law of large numbers converges to the true value of θ. These arguments are formalized and generalized as follows.

7.2 Consistency

If the true value of θ is θ_0, and the prior probability of θ_0 (or in the continuous case an arbitrary neighbourhood of θ_0) is not zero, then with increasing amounts of data x, the posterior probability that $\theta = \theta_0$ (or lies in a neighbourhood of θ_0) tends to unity. This is proved as follows:

Let $x_1, \ldots x_n$ be IID observations, each with distribution $g(x|\theta)$. Then the posterior density is

$$f(\theta|x_1, \ldots x_n) \quad \propto \quad f(\theta) \prod_{i=1}^{n} g(x_i|\theta)$$

$$\begin{aligned} &= f(\theta)\exp\left\{\sum_{i=1}^{n}\log g(x_i|\theta)\right\} \\ &= f(\theta)\exp\{n\bar{\ell}_n(\theta)\} \\ &\propto f(\theta)\exp\{n[\bar{\ell}_n(\theta)-\bar{\ell}_n(\theta_0)]\} \end{aligned}$$

say. Now, for fixed θ, $\bar{\ell}_n(\theta)-\bar{\ell}_n(\theta_0)$ is the mean of n IID random variables, and so converges in probability to its expectation

$$\int\{\log g(x|\theta)-\log g(x|\theta_0)\}g(x|\theta_0)dx.$$

This can be shown to be less than 0, except that it is equal to 0 when $\theta=\theta_0$. Thus, for $\theta\neq\theta_0$, it follows that $\exp\{n[\bar{\ell}_n(\theta,x)-\bar{\ell}_n(\theta_0,x)]\}$ tends to 0 with probability 1 as $n\to\infty$, but remains at 1 for $\theta=\theta_0$. This is sufficient, provided $f(\theta_0)\neq 0$, to prove the assertion.

Thus, as long as the prior distribution gives non–zero weight to the true value of θ, eventually, the posterior probability will concentrate on the true value.

7.3 Asymptotic Normality

When θ is continuous, this argument can be extended to obtain the approximate form of the posterior distribution when n is large. By the argument in the previous section, as n increases $\exp\{n[\bar{\ell}_n(\theta,x)-\bar{\ell}_n(\theta_0,x)]\}$ is negligibly small on all but a vanishingly small neighbourhood of θ_0. Hence $f(\theta)$ can be regarded as constant over this neighbourhood and we obtain

$$\begin{aligned} f(\theta|x_1,\ldots x_n) &\propto \exp\{n\bar{\ell}_n(\theta)\} \\ &\propto \exp\{n[\bar{\ell}_n(\theta)-\bar{\ell}_n(\hat{\theta})]\}, \end{aligned}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ. Now we expand the exponent in this expression as a Taylor series about $\hat{\theta}$. The first two terms are both zero because the expression is corrected to be zero at this point, and it has a maximum there, so

$$n[\bar{\ell}_n(\theta)-\bar{\ell}_n(\hat{\theta})] \approx \tfrac{1}{2}\,n\,\bar{\ell}''(\hat{\theta})\times(\theta-\hat{\theta})^2.$$

Then

$$f(\theta|x_1,\ldots x_n) \propto \exp\{-\tfrac{1}{2}\,n\left[-\bar{\ell}''(\hat{\theta})\right]\times(\theta-\hat{\theta})^2\}.$$

That is

$$\theta|x\sim\text{Normal}\left(\hat{\theta},\left[-n\bar{\ell}''(\hat{\theta})\right]^{-1}\right).$$

So, as $n \to \infty$, the distribution of the posterior is approximately Normal about the maximum likelihood estimate $\hat{\theta}$, and with variance given by minus the second derivative of the log likelihood at the maximum. Notice again, that this result is true independently of the prior specification, provided the prior is not zero at the true value.

This result has a number of uses. First, it can be used directly to obtain approximate posterior probabilities in situations where computations to obtain the true posterior are difficult. Second, it can give useful starting values for numerical computations where analytical solutions are intracable. Most importantly perhaps though, it demonstrates that once you get enough data, concerns about the particular prior you have selected become irrelevant. Two individuals might specify quite different forms for their prior beliefs, but eventually, once enough data become available, their posterior inferences will agree.

Example 7.1 *(Normal mean.) Let $X_1, \ldots, X_n$ be a set of independent variables from* Normal (θ, σ^2), *where σ^2 is known.*

As usual, this gives the likelihood

$$f(x|\theta) \propto \exp\{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}\}$$

Thus, we can take

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

so that

$$\frac{d\log(f(x|\theta))}{d\theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)$$

and

$$\frac{d^2\log(f(x|\theta))}{d^2\theta} = -n/\sigma^2.$$

Consequently, the m.l.e. $\hat{\theta} = \overline{x}$ and $-n\overline{\ell}''(\hat{\theta}) = n/\sigma^2$. Hence, asymptotically as $n \to \infty$,

$$\theta|x \sim \text{Normal}\,(\overline{x}, \sigma^2/n).$$

This is true for *any* prior distribution which places non–zero probability around the true value of θ.

Example 7.2 *(Binomial sample.) Consider again the likelihood model $x \sim$* Binomial (n, θ).

So

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}; \qquad x = 0, \ldots, n.$$

Thus

$$\log(f(x|\theta)) = x\log\theta + (n-x)\log(1-\theta).$$

So,

$$\frac{d\log(f(x|\theta))}{d\theta} = \frac{x}{\theta} - \frac{(n-x)}{1-\theta}$$

and

$$\frac{d^2\log l(\theta)}{d^2\theta} = \frac{-x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2}.$$

Consequently, $\hat{\theta} = x/n$ and

$$-n\bar{\ell}''(\hat{\theta}) = \frac{n\hat{\theta}}{\hat{\theta}^2} - \frac{n(1-\hat{\theta})}{(1-\hat{\theta})^2} = \frac{n}{\hat{\theta}(1-\hat{\theta})}.$$

Thus, as $n \to \infty$,

$$\theta|x \sim \text{Normal}\left(\frac{x}{n}, \frac{\frac{x}{n}(1-\frac{x}{n})}{n}\right).$$

7.4 Exercises

Exercise 7.1 *Find the asymptotic posterior distribution for θ for each of the two models in exercise 2.1.*

Exercise 7.2 *Find the asymptotic posterior distribution for b in the Pareto model of exercise 3.3.*