M249

Practical modern statistics

# Bayesian statistics

## About this module

M249 *Practical modern statistics* uses the software packages *IBM SPSS Statistics* (SPSS Inc.) and *WinBUGS*, and other software. This software is provided as part of the module, and its use is covered in the *Introduction to statistical modelling* and in the four computer books associated with *Books 1* to *4*.

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 60 90; email general-enquiries@open.ac.uk).
Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.
To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure (tel. +44 (0)1908 858779; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

*Note to reader*
Mathematical/statistical content at the Open University is usually provided to students in printed books, with PDFs of the same online. This format ensures that mathematical notation is presented accurately and clearly. The PDF of this extract thus shows the content exactly as it would be seen by an Open University student. Please note that the PDF may contain references to other parts of the module and/or to software or audio-visual components of the module. Regrettably mathematical and statistical content in PDF files is unlikely to be accessible using a screenreader, and some OpenLearn units may have PDF files that are not searchable. You may need additional help to read these documents.

The Open University, Walton Hall, Milton Keynes MK7 6AA.

First published 2007.

Printed and bound in the United Kingdom by The Charlesworth Group, Wakefield.

ISBN 978 0 7492 1369 5
1.1

# Contents

# The Bayesian approach

## Introduction

One of the key ideas in Bayesian statistics is that knowledge about anything unknown can be expressed probabilistically. For example, suppose we are interested in the proposition $E$, where

A proposition is any kind of statement which can be true or false.

$$E = \text{a major earthquake will occur in Europe in the next ten years.}$$

Since $E$ is unobserved, there is uncertainty as to whether or not $E$ is true. In Bayesian statistics, knowledge about $E$ can be expressed by $P(E)$, the probability that $E$ is true. The closer the value of $P(E)$ is to 1, the more likely $E$ is to be true, and the closer $P(E)$ is to 0, the more likely $E$ is to be false. In a Bayesian statistical analysis, probabilities such as $P(E)$, which represent knowledge about $E$, need to be estimated. How this might be done is the focus of Section 1.

Suppose that data are observed, or new information becomes available, to help learn about a proposition. The data (or information) will update the knowledge about the proposition. This updated knowledge can be expressed probabilistically. In Section 2, Bayes' theorem is introduced. This provides a method of calculating the probability representing the updated knowledge.

In most statistical problems, it is not simply the probability of a proposition that is of interest, but the more general problem of learning about an unknown parameter or, indeed, any unobserved quantity. In Bayesian statistics, knowledge about an unknown parameter, or quantity, can also be expressed probabilistically. Bayes' theorem provides a method of calculating the updated knowledge about the unknown parameter or quantity, as represented probabilistically, after observing any relevant data or information. How this can be done is discussed in Section 3, and a framework for Bayesian inference is established.

# 1 Estimating probabilities

Often in Bayesian statistics the probability of some proposition has to be estimated. In this section, several ways in which this might be done are discussed. These include simply calculating probabilities and using relative frequencies to estimate probabilities. You will be familiar with these methods from classical statistics. Probabilities that cannot be estimated objectively and need to be estimated subjectively are also considered.

## Calculating probabilities

Some probabilities can simply be calculated, as illustrated in Example 1.1.

### Example 1.1 Tossing a coin

If a coin is tossed, then assuming that the coin is fair, the probability that it lands heads up is calculated as

$$P(\text{head}) = \frac{\text{number of ways of getting a head}}{\text{number of possible outcomes (head or tail)}} = \frac{1}{2}. \quad ♦$$

### Activity 1.1 Rolling a die

A fair die with faces numbered from 1 to 6 is rolled. Calculate the probability that the die lands with an even number uppermost.

In Example 1.1 and Activity 1.1, there is little uncertainty about the values of the probabilities, and it is unlikely there would be disagreement about the values.

## Estimating probabilities using relative frequencies

Another method of estimating the probability of an event is to use the observed, or hypothetical, relative frequency of the event. The **relative frequency** of an event is the number of occurrences of the event divided by the number of times the event could have occurred. This is illustrated in Examples 1.2 to 1.4.

### Example 1.2 Probability of Down's syndrome

In every pregnancy there is a risk that the foetus has Down's syndrome. The risk increases with maternal age, so an older mother is more likely to be carrying a foetus with Down's syndrome than is a younger mother.

Several large-scale surveys have been carried out to estimate, for different ages of the mother, the probability that a foetus has Down's syndrome. For each maternal age, data are available on the proportion of foetuses with Down's syndrome. These data provide a direct estimate of the relative frequency, and hence of the probability, of Down's syndrome for different ages of the mother. Table 1.1 (overleaf) shows estimates of $P(\text{Down's})$, the probability of Down's syndrome at 40 weeks' gestation, for different maternal ages.

These data were obtained in June 2004 from the website of the Fetal Medicine Foundation http://www.fetalmedicine.com.

*Table 1.1* Estimated probability that a foetus has Down's syndrome, by maternal age

| Maternal age (years) | Estimate of $P$(Down's) |
|---|---|
| 20 | 1/1527 |
| 25 | 1/1352 |
| 30 | 1/895 |
| 31 | 1/776 |
| 32 | 1/659 |
| 33 | 1/547 |
| 34 | 1/446 |
| 35 | 1/356 |
| 36 | 1/280 |
| 37 | 1/218 |
| 38 | 1/167 |
| 39 | 1/128 |
| 40 | 1/97 |
| 41 | 1/73 |
| 42 | 1/55 |
| 43 | 1/41 |
| 44 | 1/30 |
| 45 | 1/23 |

Note that the probabilities in Table 1.1 are *estimates* and may not be equal to the true, underlying values of the probabilities. However, they are estimated from large amounts of data, so they should be very good estimates. ♦

## *Example 1.3 Probability of a volcanic eruption*

In January 1993, a conference on predicting volcanic eruptions was held at the foot of the Galeras volcano in Colombia. The highlight of the conference was to be a trip into the crater. After much debate concerning the risk (that is, the probability) of Galeras erupting, the trip went ahead. However, Galeras erupted during the trip, killing six scientists and three tourists.

Aspinall, W.P., Woo, G., Voight, B., Baxter, P.J. *et al.* (2003) Evidence-based volcanology: application to eruption crises. *Journal of Volcanology and Geothermal Research*, **128**, 273–285.

Suppose that prior to the trip into the crater, an estimate was required of the probability of an imminent eruption, where 'imminent' is taken to mean within the next seven days. How might a suitable estimate be obtained? One simple method of estimating the probability of eruption within the next week is to use data on previous eruptions to calculate the relative frequency of weeks in a year when Galeras erupted.

Before January 1993, there was approximately one eruption per year, on average. That is, out of the 52 weeks in a year there was one week, on average, in which Galeras erupted. Therefore, assuming that eruptions do not follow any temporal pattern — for example, all occurring in summer — an estimate of the relative frequency of weeks with an eruption in a year is $\frac{1}{52}$. Therefore the probability of an eruption being imminent can be estimated as $\frac{1}{52}$. ♦

Since the estimates of $P$(Down's) in Example 1.2 were based on a large amount of data, we can be fairly confident about them. However, the estimate of the probability $P$(imminent eruption) in Example 1.3 was based on far less data, and consequently we may be less confident about it. Indeed, it is possible that there could be disagreement about its value — especially from expert volcanologists who may have different knowledge about Galeras and volcanic behaviour. For example, eruptions might not be independent events, in which case the recent history of eruptions might affect the probability of an eruption in the next week.

In Example 1.4, the problem of estimating the probability that a defendant is guilty in a court case is considered. In this example, there are no data that can be used directly to calculate relative frequencies. Instead, relative frequencies need to be hypothesized.

### *Example 1.4* *Probability of guilt*

Denis John Adams was tried on a charge of sexual assault in January 1995. Suppose that before any evidence from either the prosecution or the defence had been heard, the court wished to have an estimate of the probability that Adams is guilty.

Dawid, A.P. (2007) Statistics and the law. In Tybjerg, K., Swenson-Wright, J. and Bell, A. (eds) *Evidence*. Cambridge University Press.

Before any prosecution or defence evidence is presented, it might be reasonable to assume that the culprit is a male aged between 18 and 60. There were approximately 150 000 males of this age who lived locally to where the crime took place. Of course, the culprit may not be local. This increases the hypothesized number of possible men who could be the culprit from around 150 000 to 200 000, say. Before observing any evidence, Adams is known only to be one of the possible culprits, so $P(\text{Adams guilty})$ can be estimated to be $\frac{1}{200\,000}$.

It is possible that not everyone will agree with the estimate of $\frac{1}{200\,000}$ for $P(\text{Adams guilty})$. For example, the prosecution may argue that the hypothesized number of possible men who could be the culprit is much less than 200 000, so that the estimate for $P(\text{Adams guilty})$ should in fact be larger. ♦

### *Activity 1.2* *Coronary heart disease*

Coronary heart disease (CHD) is currently the most common cause of death in the UK. In 2002, of 288 332 male deaths, 64 473 were from CHD, and of 318 463 female deaths, 53 003 were from CHD.

These data were obtained in June 2004 from the British Heart Foundation Statistics website http://www.heartstats.org.

(a) Estimate the following probabilities.

(i) The probability that a randomly chosen UK man who dies in a future year will die of CHD.

(ii) The probability that a randomly chosen UK woman who dies in a future year will die of CHD.

(b) How accurate are your estimates likely to be?

## *Estimating probabilities subjectively*

In some situations, it may not be possible to observe, or even hypothesize, the relative frequency of an event. For example, an event may occur only very rarely or be a ‘one-off’ event. This is illustrated in Examples 1.5 and 1.6.

### *Example 1.5* *Avian influenza*

In 2006, the spread of the H5N1 influenza virus in birds caused widespread concern about the likelihood of a world epidemic (or pandemic) of this deadly type of influenza in humans. This would occur if the avian H5N1 influenza virus were to mutate so as to transmit easily between humans. Based on the current understanding of the biology of the influenza virus, and the observation that human pandemics of this type of influenza have occurred in the past, scientists are agreed that the chance of such a mutation occurring at some point in the future is high. However, there are no reliable frequency data from which to estimate directly the probability that this will occur in the next year. ♦

### Example 1.6 US space missions

In January 2004, the US president George W. Bush unveiled a plan to return Americans to the moon by 2020, with the aim that this would be used as a stepping stone for a manned mission to Mars. Consider the following probabilities:

$P$(Americans return to the moon by 2020),

$P$(manned mission to Mars by 2035).

These are both probabilities of one-off events with no data available from which to estimate relative frequencies. ♦

The probabilities in Examples 1.5 and 1.6 cannot be estimated objectively. Any estimates of the probabilities will be subjective to some degree, and as such will reflect the opinions and beliefs of whoever is making the estimates. For this reason, Bayesian statistics often refers to probabilities as representing beliefs, or opinions, about a proposition. Naturally, whenever estimates are made subjectively, there may be disagreement about the estimates.

Estimating the probability of an event accurately can be difficult, especially if the person estimating the probability is not confident, or familiar, with using probabilities. To illustrate this, try estimating some probabilities yourself.

### Activity 1.3 Position of the letter r

This activity concerns the use of the letter r in English words. The probabilities $p_1$ and $p_3$ are defined as follows:

$p_1 = P$(an English word begins with the letter r),

$p_3 = P$(an English word has r as the third letter).

Which of these two probabilities would you estimate to be the larger? Explain your answer.

## Summary of Section 1

In this section, the estimation of probabilities has been discussed. The simplest way to estimate the probability of an event is to use the observed, or hypothesized, relative frequency of the event. When this is not possible, more subjective methods for obtaining probability estimates must be used. You have seen that estimating probabilities is not always easy.

## Exercise on Section 1

### Exercise 1.1 Children's lunchboxes

(a) In a study to investigate what UK children have in their packed lunches at school, 720 mothers were questioned. Of these mothers, 374 said that they packed crisps into their child's lunchbox every day. Estimate the probability that a UK schoolchild has crisps in their lunchbox every day.

(b) In another study, the contents of the lunchboxes of 28 UK children in a class were observed for one week. It was found that 19 of the children had crisps every day that week. Use data from this study to estimate the probability that a UK schoolchild has crisps in their lunchbox every day.

(c) Which of the estimates that you calculated in parts (a) and (b) do you think is the more reliable? Explain your answer.

# 2 Bayes' theorem

In this section, a result which is at the heart of Bayesian statistics is introduced. This result is known as Bayes' theorem. One of the key elements of Bayes' theorem is conditional probability; this is discussed in Subsection 2.1. Bayes' theorem is introduced in Subsection 2.2. You will see how it can be used to update beliefs about a proposition when data are observed or information becomes available. In fact, Bayes' theorem can be used repeatedly to update beliefs as additional data or information become available. This is illustrated in Subsection 2.3.

## 2.1 Conditional probability

Conditional probabilities were discussed briefly in the *Introduction to statistical modelling*. Since they are very important in Bayesian statistics, they will be considered further in this subsection. The ideas of conditional probability and independence are reviewed in Example 2.1.

### *Example 2.1* Heart disease and gender

The data on deaths from coronary heart disease (CHD) by gender, which were introduced in Activity 1.2, are summarized in Table 2.1.

*Table 2.1* UK deaths in 2002 from coronary heart disease (CHD) by gender

| | Gender | | |
|---|---|---|---|
| Died from CHD | Male | Female | Total |
| Yes | 64 473 | 53 003 | 117 476 |
| No | 223 859 | 265 460 | 489 319 |
| Total | 288 332 | 318 463 | 606 795 |

Let the variable $X$ represent the gender of a person who died, and let the variable $Y$ indicate whether or not a person died from CHD. An estimate of $P(Y = \text{yes})$, the probability that a randomly chosen individual died from CHD, is given by

$$\frac{\text{number who died from CHD}}{\text{total number who died}} = \frac{117\,476}{606\,795} \simeq 0.1936.$$

Now suppose that the additional information that the person who died was male is available — that is, $X$ takes the value 'male'. With this extra knowledge, $P(Y = \text{yes})$ is no longer the appropriate probability that death was due to CHD. The probability required can be estimated from the data in Table 2.1 on males who died:

You estimated this probability in Activity 1.2.

$$\frac{\text{number of males who died from CHD}}{\text{number of males who died}} = \frac{64\,473}{288\,332} \simeq 0.2236.$$

The probability that a person died from CHD (that is, $Y$ takes the value 'yes'), when it is known that the person is male ($X$ takes the value 'male') is the conditional probability $P(Y = \text{yes}|X = \text{male})$.

Recall that two random variables $X$ and $Y$ are independent if, for all values of $x$ and $y$,

$$P(Y = y|X = x) = P(Y = y).$$

In this case, the estimated conditional probability that $Y = \text{yes}$, given $X = \text{male}$, is different from the estimated probability that $Y = \text{yes}$; that is,

$$P(Y = \text{yes}|X = \text{male}) \neq P(Y = \text{yes}).$$

Therefore $X$ and $Y$ may be dependent variables. ♦

### *Activity 2.1 Conditional probabilities of Down's syndrome*

Estimates of $P(\text{Down's})$, the probability that a foetus has Down's syndrome for different ages of the mother, were given in Table 1.1. The probabilities, which are reproduced in Table 2.2, are in fact the conditional probabilities $P(\text{Down's}|\text{mother's age})$.

(a) Use Table 2.2 to write down the following conditional probabilities.

(i) $P(\text{Down's}|\text{mother's age} = 25 \text{ years})$

(ii) $P(\text{Down's}|\text{mother's age} = 40 \text{ years})$

(b) Is it possible to conclude from your answers to part (a) that more foetuses with Down's syndrome occur in pregnant women aged 40 than in pregnant women aged 25?

*Table 2.2* Estimated probability that a foetus has Down's syndrome, by maternal age

| Maternal age (years) | Estimate of $P(\text{Down's})$ |
|---|---|
| 20 | 1/1527 |
| 25 | 1/1352 |
| 30 | 1/895 |
| 31 | 1/776 |
| 32 | 1/659 |
| 33 | 1/547 |
| 34 | 1/446 |
| 35 | 1/356 |
| 36 | 1/280 |
| 37 | 1/218 |
| 38 | 1/167 |
| 39 | 1/128 |
| 40 | 1/97 |
| 41 | 1/73 |
| 42 | 1/55 |
| 43 | 1/41 |
| 44 | 1/30 |
| 45 | 1/23 |

In general, if the probability that an event $A$ occurs depends on whether or not an event $B$ has occurred, then we speak of the **conditional probability of *A* given *B***. The conditional probability is defined in the following box.

**Conditional probability**

For two events $A$ and $B$, the **conditional probability** of $A$ given $B$ is denoted $P(A|B)$, and is defined by the formula

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}. \qquad (2.1)$$

The probability $P(A \text{ and } B)$ is the **joint probability** of $A$ and $B$.

In Example 2.1, conditional probabilities were estimated from a contingency table. More generally, conditional probabilities can be calculated using Formula (2.1). This is illustrated in Example 2.2.

### *Example 2.2 Calculating a conditional probability*

Data on deaths from CHD by gender were given in Example 2.1 (see Table 2.1). The variable $X$ represents the gender of a person who died, and the variable $Y$ indicates whether or not a person died from CHD. In Example 2.1, the conditional probability $P(Y = \text{yes}|X = \text{male})$ was estimated directly to be 0.2236. This probability could have been calculated using Formula (2.1), as follows.

First, estimates of the probabilities $P(Y = \text{yes and } X = \text{male})$ and $P(X = \text{male})$ are required. Using the data in Table 2.1,

$$P(X = \text{male}) = \frac{\text{number of male deaths}}{\text{total number of deaths}} = \frac{288\,332}{606\,795} \simeq 0.4752,$$

$$\begin{aligned} P(Y = \text{yes and } X = \text{male}) &= \frac{\text{number of men who died from CHD}}{\text{total number of deaths}} \\ &= \frac{64\,473}{606\,795} \simeq 0.1063. \end{aligned}$$

Then Formula (2.1) gives

$$P(Y = \text{yes}|X = \text{male}) = \frac{P(Y = \text{yes and } X = \text{male})}{P(X = \text{male})} \simeq \frac{0.1063}{0.4752} \simeq 0.2237.$$

The slight discrepancy between this result and that obtained directly in Example 2.1 is due to rounding error. ♦

### *Activity 2.2 Calculating another conditional probability*

(a) Use the data in Table 2.1 to estimate the conditional probability $P(X = \text{male}|Y = \text{yes})$.

(b) The joint probability $P(Y = \text{yes and } X = \text{male})$ was estimated to be 0.1063 in Example 2.2. In Example 2.1, the probability $P(Y = \text{yes})$ was estimated to be 0.1936. Use Formula (2.1) to estimate the conditional probability $P(X = \text{male}|Y = \text{yes})$. Check that this estimate of the conditional probability is the same as that obtained in part (a).

Using Formula (2.1), for any events $A$ and $B$,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)},$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

Hence $P(A|B) = P(B|A)$ only when $P(A) = P(B)$. In general, for events $A$ and $B$,

$$P(A|B) \neq P(B|A).$$

The ordering of the conditioning is important. An illustration of this is given by Example 2.2 and Activity 2.2: in Example 2.2, you saw that $P(Y = \text{yes}|X = \text{male}) \simeq 0.2237$, whereas in Activity 2.2, you found that $P(X = \text{male}|Y = \text{yes}) \simeq 0.5491$.

In some areas of application, misinterpreting the ordering of the conditioning can have serious consequences. One such area is in court cases. In a court case, a juror will vote for a guilty verdict only if he or she believes 'beyond reasonable doubt' that the defendant is guilty — that is, if, after hearing all the evidence, he or she assesses the probability that the defendant is guilty to be close to 1. The jurors therefore need to estimate the conditional probability $P(\text{defendant guilty}|\text{evidence})$. However, as you will see in Example 2.3, it can be all too easy for a juror to interpret $P(\text{evidence}|\text{defendant guilty})$ as being the same as $P(\text{defendant guilty}|\text{evidence})$. This is known as the 'prosecutor's fallacy' as it can lead to a juror believing that the probability that the defendant is guilty is larger than it actually is.

### *Example 2.3 The prosecutor's fallacy*

The case of Denis John Adams, who was tried on the charge of sexual assault in January 1995, was introduced in Example 1.4. The prosecution case rested on forensic evidence, which will be called $M$:

$M$ = DNA match between Adams and a sample, accepted as
being from the culprit, taken from the victim.

The prosecutor's forensic expert testified that for a randomly chosen person, the probability that their DNA matched that of the sample was 1/200 000 000. The defence argued that this probability was in fact 1/2 000 000. It is tempting to assign the value 1/200 000 000 (or 1/2 000 000) to the conditional probability $P(\text{Adams not guilty}|M)$. If this is done then, since this probability is very small — 1/200 000 000 according to the prosecution, 1/2 000 000 according to the defence — and given that there is a DNA match, it would be reasonable to conclude that Adams must be guilty. In fact, the probabilities of a DNA match provided by the prosecution and defence were estimates of the conditional probability $P(M|\text{Adams not guilty})$, which is not the same as $P(\text{Adams not guilty}|M)$. Basing a verdict on the probability $P(M|\text{Adams not guilty})$ is incorrect: it is an instance of the prosecutor's fallacy. ♦

## 2.2 Bayes' theorem

Formula (2.1) can be rearranged to give a formula for the joint probability of $A$ and $B$ in terms of $P(B)$ and $P(A|B)$:

$$P(A \text{ and } B) = P(A|B) \times P(B). \tag{2.2}$$

However, the order of $A$ and $B$ in Formula (2.1) is arbitrary, so

$$P(B \text{ and } A) = P(B|A) \times P(A).$$

Since $P(A \text{ and } B) = P(B \text{ and } A)$, it follows that

$$P(A|B) \times P(B) = P(B|A) \times P(A).$$

Provided that $P(B) \neq 0$, dividing by $P(B)$ leads to an alternative formula for calculating a conditional probability:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}. \tag{2.3}$$

The use of this formula to calculate a conditional probability is illustrated in Example 2.4.

### *Example 2.4* *Using the alternative formula*

Formula (2.3) can be used to obtain an estimate of the probability that a person who is known to have died of CHD is male, as follows:

$$P(X = \text{male}|Y = \text{yes}) = \frac{P(Y = \text{yes}|X = \text{male}) \times P(X = \text{male})}{P(Y = \text{yes})}.$$

In Example 2.1, $P(Y = \text{yes}|X = \text{male})$ was estimated to be 0.2236, and $P(Y = \text{yes})$ was estimated to be 0.1936. In Example 2.2, $P(X = \text{male})$ was estimated to be 0.4752. Therefore

$$P(X = \text{male}|Y = \text{yes}) \simeq \frac{0.2236 \times 0.4752}{0.1936} \simeq 0.5488.$$

This is the same as the estimate that you obtained in part (a) of Activity 2.2. ♦

You may be wondering why Formula (2.3) might be preferred to Formula (2.1) for calculating a conditional probability. The reason is that it is often easier to calculate $P(A \text{ and } B)$ using $P(B|A) \times P(A)$ than it is to calculate it directly. This is illustrated in Example 2.5.

### *Example 2.5* *Probability of guilt given prosecution's evidence*

In Example 2.3, you saw that the prosecution case against Denis John Adams rested on forensic evidence $M$ of a DNA match. The judge and jury require an estimate of the probability that Adams is guilty, given the evidence $M$. Using Formula (2.1) for a conditional probability, this is given by

$$P(\text{Adams guilty}|M) = \frac{P(\text{Adams guilty and } M)}{P(M)}.$$

However, thinking about the joint probability $P(\text{Adams guilty and } M)$ is not easy! On the other hand, using Formula (2.3) for a conditional probability gives

$$P(\text{Adams guilty}|M) = \frac{P(M|\text{Adams guilty}) \times P(\text{Adams guilty})}{P(M)}.$$

First consider $P(M|\text{Adams guilty})$. If Adams is guilty, then the probability of a DNA match might be assumed to be very close to 1. (It might not be exactly 1 due to the possibility of errors in the DNA test.) For simplicity, set

$$P(M|\text{Adams guilty}) = 1.$$

If $P(M|\text{Adams guilty}) = 1$, then $P(M \text{ and Adams guilty})$ is equal to $P(\text{Adams guilty})$. In reality, $P(M|\text{Adams guilty})$ is not exactly 1, so these probabilities will not be exactly the same.

In Example 1.4, $P(\text{Adams guilty})$ was estimated to be $1/200\,000$. Therefore

$$P(\text{Adams guilty}|M) \simeq \frac{1 \times \frac{1}{200\,000}}{P(M)}.$$

The problem of how to calculate $P(M)$ remains. The solution is to use another general result involving conditional probabilities. This is discussed below. ♦

Calculating $P(B)$ in the denominator of Formula (2.3) can be tricky. However, conditional probabilities can be used to simplify its calculation, as follows.

When $B$ occurs, either $A$ occurs, or $A$ does not occur. Therefore

$$P(B) = P(B \text{ and } A) + P(B \text{ and not } A).$$

Using Formula (2.2), this becomes

$$P(B) = P(B|A) \times P(A) + P(B|\text{not } A) \times P(\text{not } A). \tag{2.4}$$

This is the result needed. Activity 2.3 will give you some practice at using this formula.

### Activity 2.3 Calculating the probability of CHD

Data concerning deaths from coronary heart disease and gender were given in Table 2.1.

(a) Use the data to estimate the conditional probability $P(Y = \text{yes}|X = \text{female})$.

(b) In Example 2.1, $P(Y = \text{yes}|X = \text{male})$ was estimated to be 0.2236, and in Example 2.2, $P(X = \text{male})$ was estimated to be 0.4752. Use these estimates and Formula (2.4) to calculate $P(Y = \text{yes})$.

### Example 2.6 Probability of guilt given prosecution's evidence, continued

Formula (2.4) can be used to complete the calculation of the probability that Adams is guilty given evidence of a DNA match, which was begun in Example 2.5.

In that example, you saw that

$$\begin{aligned} P(\text{Adams guilty}|M) &= \frac{P(M|\text{Adams guilty}) \times P(\text{Adams guilty})}{P(M)} \\ &\simeq \frac{1 \times \frac{1}{200\,000}}{P(M)}. \end{aligned} \tag{2.5}$$

Using Formula (2.4) to calculate $P(M)$ gives

$$\begin{aligned} P(M) = {} & P(M|\text{Adams guilty}) \times P(\text{Adams guilty}) \\ & + P(M|\text{Adams not guilty}) \times P(\text{Adams not guilty}). \end{aligned} \tag{2.6}$$

In Example 2.5, $P(M|\text{Adams guilty})$ was set equal to 1, and the estimate used for $P(\text{Adams guilty})$ was 1/200 000. Also,

$$P(\text{Adams not guilty}) = 1 - P(\text{Adams guilty}) \simeq 1 - \frac{1}{200\,000} = \frac{199\,999}{200\,000}.$$

Two estimates of $P(M|\text{Adams not guilty})$ were given in Example 2.3 — the prosecution's estimate (1/200 000 000) and the defence's estimate (1/2 000 000). Using the defence's estimate in (2.6) gives

$$\begin{aligned} P(M) &\simeq 1 \times \frac{1}{200\,000} + \frac{1}{2\,000\,000} \times \frac{199\,999}{200\,000} \\ &= \frac{2\,000\,000}{2\,000\,000 \times 200\,000} + \frac{199\,999}{2\,000\,000 \times 200\,000} \\ &= \frac{2\,199\,999}{2\,000\,000 \times 200\,000}. \end{aligned}$$

Substituting this in (2.5) gives

$$\begin{aligned} P(\text{Adams guilty}|M) &\simeq \left(1 \times \frac{1}{200\,000}\right) \bigg/ \frac{2\,199\,999}{2\,000\,000 \times 200\,000} \\ &= \frac{2\,000\,000}{2\,199\,999} \\ &\simeq 0.91. \end{aligned}$$

Note that although the probability of guilt given the prosecution's evidence $M$ is high, it is not as high as that given by the prosecutor's fallacy — that is, $1 - 1/2\,000\,000 = 0.999\,999\,5$.

So far, only the prosecution's evidence $M$ has been considered. None of the defence's evidence has been taken into account. That will be done in Subsection 2.3. ♦

### *Activity 2.4 Probability of guilt according to the prosecution*

Use the method of Example 2.6 to calculate the probability that Adams is guilty given the prosecution's evidence $M$, using the prosecution's estimate of 1/200 000 000 for $P(M|\text{Adams not guilty})$. With the prosecution's estimate, does Adams seem to be guilty 'beyond all reasonable doubt'?

Formulas (2.3) and (2.4) together comprise Bayes' theorem. As its name implies, this theorem is central to Bayesian statistics. It is stated formally in the following box.

**Bayes' theorem**

For two events $A$ and $B$, provided that $P(B) \neq 0$,

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)},$$

where

$$P(B) = P(B|A) \times P(A) + P(B|\text{not } A) \times P(\text{not } A).$$

Bayes' theorem gives a method of revising probability estimates as additional information becomes available. The additional information is the information that is being conditioned on. The probability before additional information becomes available is referred to as the **prior probability**, and the revised probability using the additional information is called the **posterior probability**.

It may be helpful to think of Bayes' theorem using the diagram in Figure 2.1.

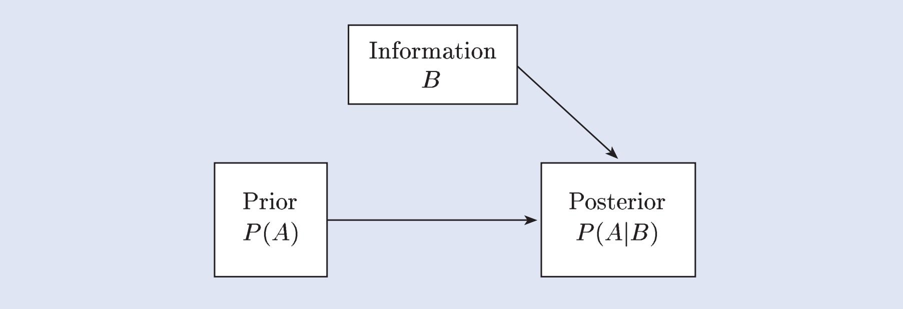


*Figure 2.1* Bayes' theorem in practice

In the case of Adams' trial, the prior probability is $P(\text{Adams guilty})$, the initial estimate of the probability that Adams is guilty before any evidence is considered; the additional information (considered so far) is the prosecution's evidence $M$; and the posterior probability is $P(\text{Adams guilty}|M)$.

## 2.3 Sequential updating

Suppose that after observing information $B$, the posterior probability $P(A|B)$ has been calculated but now some additional information $C$ is available. Bayes' theorem can be used to further revise the probability estimate. In this case, $P(A|B)$ becomes the *prior* probability, as this is the estimate *before* observing $C$. The posterior probability is $P(A|C,B)$. This can be calculated using Bayes' theorem:

$$P(A|C,B) = \frac{P(C|A,B) \times P(A|B)}{P(C|B)}, \tag{2.7}$$

where

$$P(C|B) = P(C|A,B) \times P(A|B) + P(C|\text{not } A,B) \times P(\text{not } A|B). \tag{2.8}$$

Notice that all the probabilities in (2.7) and (2.8) are conditional on $B$, since information $B$ is now part of the prior information. It may be helpful to think of sequential updating in a Bayesian analysis using the diagram in Figure 2.2.

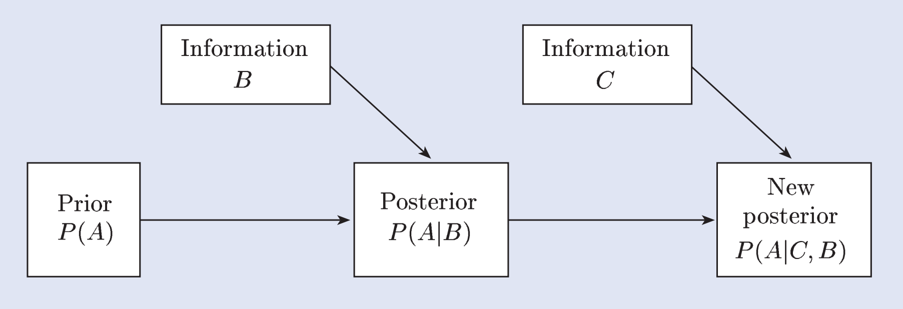


*Figure 2.2* Sequential updating

Bayes' theorem can be used each time some additional information becomes available: whenever new data become available, the current posterior probability becomes the new prior probability.

The posterior probability that Adams is guilty, given the prosecution's evidence $M$, was calculated in Example 2.6. In Example 2.7, the defence's evidence is also considered, and the new posterior probability that Adams is guilty, given evidence from both the prosecution and the defence, is calculated.

### Example 2.7 Probability of guilt considering all the evidence

In Example 2.6, the posterior probability that Adams is guilty, given the prosecution's evidence $M$, was calculated to be 0.91; that is, $P(\text{Adams guilty}|M) = 0.91$. This used the defence's prior estimate of $P(M|\text{Adams not guilty})$, which was $1/2\,000\,000$.

The defence case rested on two pieces of evidence, $E_1$ and $E_2$, where

$E_1$ = victim said Adams did not look like her assailant,

$E_2$ = Adams' girlfriend provided Adams with an alibi.

All items of evidence were considered to be independent of each other.

First, consider evidence $E_1$ — that the victim said her assailant did not look like Adams. The probability $P(\text{Adams guilty}|E_1 \text{ and } M)$ is required. The probability $P(\text{Adams guilty}|M)$ represents belief about Adams' guilt before considering evidence $E_1$, so this is the new prior. Bayes' theorem can be used to update the probability of guilt in the light of evidence $E_1$, as follows. Using (2.7) and (2.8) with all probabilities conditional on $M$,

$$\begin{aligned}&P(\text{Adams guilty}|E_1 \text{ and } M)\\&\quad= \frac{P(E_1|\text{Adams guilty}, M) \times P(\text{Adams guilty}|M)}{P(E_1|M)},\end{aligned} \tag{2.9}$$

where

$$\begin{aligned}P(E_1|M) &= P(E_1|\text{Adams guilty}, M) \times P(\text{Adams guilty}|M)\\&\quad + P(E_1|\text{Adams not guilty}, M) \times P(\text{Adams not guilty}|M).\end{aligned} \tag{2.10}$$

Items of evidence $M$ and $E_1$ are considered to be independent of each other, so the DNA evidence $M$ does not change the probability of $E_1$. Hence

$$P(E_1|\text{Adams guilty}, M) = P(E_1|\text{Adams guilty}),$$
$$P(E_1|\text{Adams not guilty}, M) = P(E_1|\text{Adams not guilty}).$$

If Adams really is guilty, the probability of $E_1$ would be low; it was thought to be about 0.1. On the other hand, if Adams is not guilty, the probability of $E_1$ would be high; it was thought to be around 0.9.

Note that it is simply a coincidence that these probabilities sum to 1.

Substituting these values in (2.10) gives

$$P(E_1|M) \simeq 0.1 \times 0.91 + 0.9 \times (1 - 0.91) = 0.172.$$

Then (2.9) becomes

$$P(\text{Adams guilty}|E_1 \text{ and } M) \simeq \frac{0.1 \times 0.91}{0.172} \simeq 0.53.$$

In the light of evidence $E_1$, the probability of Adams' guilt has dropped dramatically.

There is one final piece of evidence to consider — $E_2$, the fact that Adams' girlfriend provided him with an alibi. Taking this evidence into account, the probability required is

$$P(\text{Adams guilty}|E_2, E_1 \text{ and } M).$$

The probability $P(\text{Adams guilty}|E_1 \text{ and } M)$ becomes the new prior. The probability of guilt can be updated in the light of $E_2$ using Bayes' theorem:

$$\begin{aligned}&P(\text{Adams guilty}|E_2, E_1, M)\\&\quad= \frac{P(E_2|\text{Adams guilty}, E_1, M) \times P(\text{Adams guilty}|E_1, M)}{P(E_2|E_1, M)},\end{aligned} \tag{2.11}$$

where

$$\begin{aligned}&P(E_2|E_1, M)\\&= P(E_2|\text{Adams guilty}, E_1, M) \times P(\text{Adams guilty}|E_1, M)\\&\quad + P(E_2|\text{Adams not guilty}, E_1, M) \times P(\text{Adams not guilty}|E_1, M).\end{aligned} \tag{2.12}$$

Since all items of evidence were considered to be independent of each other,

$$P(E_2|\text{Adams guilty}, E_1, M) = P(E_2|\text{Adams guilty}),$$
$$P(E_2|\text{Adams not guilty}, E_1, M) = P(E_2|\text{Adams not guilty}).$$

The probability of $E_2$ if Adams is guilty was thought to be about 0.25, and the probability of $E_2$ if Adams is not guilty was thought to be about 0.5. Substituting these values in (2.12) gives

$$P(E_2|E_1, M) \simeq 0.25 \times 0.53 + 0.5 \times (1 - 0.53) = 0.3675,$$

and hence (2.11) becomes

$$P(\text{Adams guilty}|E_2, E_1, M) \simeq \frac{0.25 \times 0.53}{0.3675} \simeq 0.36.$$

Therefore, after all the evidence presented by both the prosecution and the defence has been considered, the probability that Adams is in fact guilty is only 0.36 — certainly not 'beyond reasonable doubt'! ♦

### Activity 2.5 Probability of guilt using prosecution's estimate

The final posterior probability $P(\text{Adams guilty}|E_2, E_1, M)$ depends on the value attached to $P(M|\text{Adams not guilty})$. The prosecution and defence did not agree on this value: the prosecution claimed that it is 1/200 000 000, while the defence claimed it should be 1/2 000 000. In Example 2.7, the value of 0.36 for the posterior probability that Adams is guilty was obtained using the defence's value of 1/2 000 000.

See Example 2.3.

In Activity 2.4, you calculated $P(\text{Adams guilty}|M)$ to be 0.999 using the prosecution's value of 1/200 000 000 for $P(M|\text{Adams not guilty})$. Use this result to calculate the posterior probability $P(\text{Adams guilty}|E_2, E_1, M)$. Is there now evidence of guilt 'beyond reasonable doubt'?

## Summary of Section 2

In this section, conditional probabilities have been reviewed and Bayes' theorem has been introduced. When Bayes' theorem is used to calculate the conditional probability $P(A|B)$, the probability of event $A$ given information $B$, the probability $P(A)$ is referred to as the prior probability, and the revised probability $P(A|B)$ as the posterior probability. You have seen that when new information $C$ becomes available, Bayes' theorem can be used again, with $P(A|B)$ as the new prior probability, to find the new posterior probability $P(A|C, B)$.

## Exercises on Section 2

### Exercise 2.1 False positive HIV tests

HIV tests are not 100% accurate and they can produce a positive result for someone who is not infected with HIV. This is known as a false positive result. Although the probability of a false positive result is small, it can affect how the results of the test should be interpreted, especially for people who have a low risk of contracting HIV.

Tests can also produce a negative result for someone who is infected with HIV (known as a false negative result).

For simplicity, suppose that an HIV test gives a positive result for all people infected with HIV, so that it does not produce any false negative results. However, suppose that it does produce some false positive results, so that the test gives a positive result to 0.005% of people who are not infected with HIV.

(a) Calculate the following probabilities.

(i) The probability that a person who is infected with HIV will have a positive test result.

(ii) The probability that a person who is not infected with HIV will have a positive test result.

(b) Suppose that a woman in a low-risk group for contracting HIV takes the test. Only 10 in 100 000 women in this group are infected with HIV.

(i) Before taking the HIV test, what is the prior probability that the woman is infected with HIV?

(ii) Given that the woman has a positive test result, calculate the posterior probability that she is infected with HIV. Would you conclude that the woman is infected with HIV?

## Exercise 2.2 Searching for a wrecked ship

In May 1968, the US nuclear submarine USS Scorpion (SSN-589) failed to arrive as expected at Norfolk, Virginia, USA, and was presumed wrecked. A team of mathematical consultants was used during the subsequent search. The sea was divided into grid squares, and a number was assigned to each square representing the probability that the wreck was in the square. The grid square with the highest probability of containing the wreck was then searched first. Although the wreck was not found in the first square searched, it was still possible that the wreck was in the square. The posterior probability that the wreck was in the square, given that it was not found, was calculated using Bayes' theorem (and the other grid square probabilities were updated accordingly, so that all the probabilities added up to 1). The second square to be searched was the one which now had the highest probability of containing the wreck. This process of searching squares and updating probabilities continued until the wreck was eventually found (in October 1968).

Information regarding USS Scorpion (SSN-589) was taken from the website of Wikipedia, The Free Encyclopedia, http://en.wikipedia.org in August 2006.

Suppose that the prior probability that the wreck is in a particular grid square is 0.4, and that the probability of finding the wreck in this grid square if it is there is 0.75. Given that the wreck is not found in this grid square, calculate the posterior probability that the wreck is in the grid square.

## Exercise 2.3 Haemophilia

Haemophilia is an inherited disease which affects only males. Although females are not affected by haemophilia, they can be carriers of the disease. If a female carrier has a son, the probability that the son is affected by haemophilia is 0.5. This probability is the same for each son a female carrier may have. If a female is not a carrier, then any son she may have will not be affected by haemophilia.

Suppose that it is not known whether or not a particular woman, called Kate, is a carrier. Suppose further that there is a history of haemophilia in her family, and that the probability that Kate is a carrier is 0.5.

(a) Kate has a son, Jack, who is not affected by haemophilia. Calculate the following probabilities.

(i) $P(\text{Jack is not affected}|\text{Kate is a carrier})$

(ii) $P(\text{Jack is not affected}|\text{Kate is not a carrier})$

(b) Given that Jack is not affected by haemophilia, calculate the posterior probability that Kate is a carrier.

(c) You are now told that Kate has a younger son, Luke, who is also unaffected by haemophilia. Using your result from part (b), calculate the new posterior probability that Kate is a carrier.

# 3 A framework for Bayesian inference

In this section, the main ideas of the Bayesian inference process are presented.

In Sections 1 and 2, beliefs about a proposition $A$ were represented probabilistically through the prior probability $P(A)$. After additional information $B$ became available, Bayes' theorem was used to update beliefs about $A$, as expressed by the posterior probability $P(A|B)$. In most statistical problems, interest focuses not on a proposition but on an unknown parameter $\theta$. In Bayesian statistics, beliefs about any unknown parameter are also represented probabilistically. Initially, this is through what is known as the **prior distribution**. Prior distributions are the subject of Subsection 3.1.

In Subsection 3.2, conditional distributions are discussed. These are required for describing the Bayesian process of inference.

Additional information which may update beliefs about $\theta$ are usually in the form of observed data $x_1, x_2, \ldots, x_n$. The information regarding $\theta$ contained in the data is represented by the **likelihood function**. The likelihood function is introduced in Subsection 3.3.

In Section 2, Bayes' theorem was used to update beliefs about a proposition after additional information became available. Bayes' theorem can also be used to update beliefs about a parameter $\theta$ after data are observed. The updated beliefs are represented by the **posterior distribution**. The posterior distribution, which summarizes all the information available about $\theta$ after observing data, is the primary focus of Bayesian inference. The posterior distribution and how it can be calculated using Bayes' theorem is discussed in Subsection 3.4.

## 3.1 Prior distributions

In Section 1, you saw how beliefs about a proposition $A$ are represented by the probability $P(A)$, which can be estimated subjectively. Beliefs about an unknown parameter $\theta$ are also represented probabilistically in Bayesian statistics. A subjective estimate can be made of the probability that the value of $\theta$ is $\theta_1$, say — that is, of the probability $P(\theta = \theta_1)$, for some value $\theta_1$.

If you are certain that $\theta = \theta_1$, then $P(\theta = \theta_1) = 1$. However, the value of $\theta$ is rarely known with certainty. Instead, there will be other values of $\theta$ that are possible. Usually, the possible values of $\theta$ are all values in some continuous interval. For example, if $\theta$ is a proportion, then the true value of $\theta$ could potentially be any value in the interval $[0, 1]$. However, for simplicity, first suppose that $\theta$ can only be one of a set of discrete values $\theta_1, \theta_2, \ldots, \theta_n$. For each possible value $\theta_i$, the probability $P(\theta = \theta_i)$ can be estimated subjectively, so that $P(\theta = \theta_i)$ represents beliefs about whether or not $\theta = \theta_i$. If $P(\theta = \theta_i)$ is estimated for all possible values of $\theta_i$, then these probabilities will form a probability distribution for $\theta$. This probability distribution gives a probabilistic representation of all the available knowledge about the parameter $\theta$, and is known as the **prior distribution**, or simply the **prior**.

Note that, although $\theta$ has a probability distribution, $\theta$ does not vary: it is not a random variable. As in classical statistics, the true value of $\theta$ is fixed but unknown. The probability distribution for $\theta$ represents *beliefs* about the true value of $\theta$.

Example 3.1 gives a simple illustration of how beliefs about a parameter $\theta$ can be used to estimate individual probabilities $P(\theta = \theta_i)$, and hence a prior for $\theta$.

### Example 3.1 Cancer at Slater School

Slater School is an elementary school in California. Staff were concerned that there was a high number of cancers among staff at the school. They thought this might be due to high-voltage power cables running past the school.

Lavine, M. (1999) What is Bayesian statistics and why everything else is wrong. *Journal of Undergraduate Mathematics and its Applications*, **20**, 165–174.

It was estimated that, nationally, the probability of an individual developing cancer was 0.03. Let the parameter $\theta$ be the probability of developing cancer at Slater School. It is assumed that this probability is the same for each of the 145 members of staff at the school, and that staff members develop cancer independently. Since $\theta$ is a probability, all values between 0 and 1 are possible values of $\theta$. However, for the sake of illustration, suppose that only four values of $\theta$ are possible: 0.03, 0.04, 0.05 and 0.06. If the value of $\theta$ is 0.03, this would mean that the cancer rate is the same at Slater School as it is nationally; the other three values (0.04, 0.05 and 0.06) represent a higher incidence of cancer than the national average.

There have been other studies investigating whether proximity to high-voltage transmission lines can cause cancer, but the results are inconclusive. Suppose that equal weight can be attached to both sides of the argument. Then it is equally likely that proximity to high-voltage power cables does not cause cancer — in which case $\theta = 0.03$, the same as the national rate — and that proximity to high-voltage transmission lines does cause cancer — so that $\theta > 0.03$. This information can be represented probabilistically as follows:

$$P(\theta = 0.03) = \tfrac{1}{2}, \quad P(\theta = 0.04 \text{ or } 0.05 \text{ or } 0.06) = \tfrac{1}{2}.$$

Suppose that there is no information to suggest that any of the values 0.04, 0.05 and 0.06 is more likely than any other. This can be expressed as

$$P(\theta = 0.04) = P(\theta = 0.05) = P(\theta = 0.06) = \tfrac{1}{6}.$$

A prior distribution has been defined for $\theta$. The probability mass function $p(\theta)$ of this prior is shown in Table 3.1. ♦

*Table 3.1* The p.m.f. of a prior distribution for $\theta$

| $\theta$ | 0.03 | 0.04 | 0.05 | 0.06 |
|---|---|---|---|---|
| $p(\theta)$ | $\frac{1}{2}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

### Activity 3.1 Cancer at Slater School

In Example 3.1, concerning the incidence of cancer at Slater School, what is the most likely value of $\theta$ based on the prior beliefs about $\theta$?

In most cases, a parameter $\theta$ will take values over an interval (such as $\theta > 0$ or $-\infty < \theta < \infty$). So instead of having a probability mass function $p(\theta)$ defined at a discrete set of values $\theta_1, \dots, \theta_n$, the prior distribution will have a probability density function $f(\theta)$, defined for all possible values of $\theta$. This is called the **prior density**. From now on, unless specified otherwise, it will be assumed that the prior for $\theta$ is a continuous probability distribution.

In Example 3.1, when considering a small number of possible values of $\theta$, it was possible to estimate the probability that $\theta$ takes each individual possible value. However, this cannot be done when $\theta$ can take values defined on an interval. Instead, a prior density must be estimated whose shape represents beliefs about $\theta$. The general idea of how this might be done is illustrated in Example 3.2.

### Example 3.2 Are politicians truthful?

*Guardian*, 23 March 2004.

In an opinion poll, a number of adults in Great Britain were asked whether or not they trust politicians to tell the truth. Let $\theta$ be the unknown proportion of individuals who trust politicians to tell the truth; since $\theta$ is a proportion, it takes values between 0 and 1. Therefore, to represent prior beliefs about $\theta$, a probability density function $f(\theta)$ must be defined for $\theta$ in the interval $[0, 1]$.

Suppose that you believe that $\theta$ is unlikely to be smaller than 0.05, or bigger than 0.9, and that you believe the most likely value of $\theta$ is 0.4. A sketch of a possible prior density representing these beliefs about $\theta$ is given in Figure 3.1. Notice that the density lies almost entirely between the values 0.05 and 0.9, because it is thought that $\theta$ is unlikely to be smaller than 0.05 or larger than 0.9. Also, the prior density peaks at the value $\theta = 0.4$, as 0.4 is believed to be the most likely value of $\theta$.

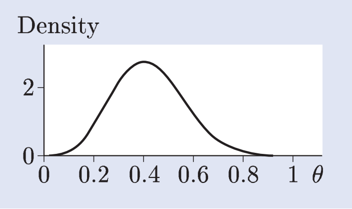


*Figure 3.1* A possible prior density for $\theta$

Now suppose that you believe that the most likely value of $\theta$ is 0.4, but that $\theta$ is unlikely to be smaller than 0.3 or larger than 0.5. A sketch of a second possible prior density is shown in Figure 3.2. The prior densities in Figures 3.1 and 3.2 both have a peak at $\theta = 0.4$, as this is thought to be the most likely value of $\theta$ in both cases. However, since the area under the graph of $f(\theta)$ is 1 for both densities (because $f(\theta)$ is a probability density function), and the range of possible values of $\theta$ is smaller for the density in Figure 3.2, this prior density is narrower and taller than the prior density in Figure 3.1. The narrower prior density in Figure 3.2 represents a stronger belief that $\theta$ is close to 0.4 than that represented by the prior density in Figure 3.1. ♦

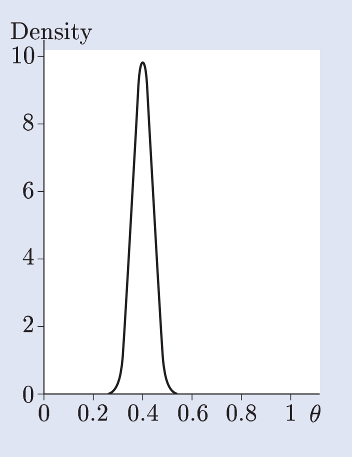


*Figure 3.2* Another possible prior density for $\theta$

In general, the more uncertain you are about the value of $\theta$, the flatter and wider your prior density should be. Conversely, the more confident you are about the value of $\theta$, the taller and narrower your prior density should be. If you are uncertain about the value of $\theta$, and hence have chosen a prior density that is quite flat, then this prior is said to be **weak**. On the other hand, if you are fairly confident about the value of $\theta$ (so that your prior density is tall and narrow), then this prior is said to be **strong**. Thus the prior in Figure 3.2 is stronger than the prior in Figure 3.1.

### Activity 3.2 Sketching prior densities

In Example 3.2, $\theta$ is the proportion of individuals in Great Britain who trust politicians to tell the truth.

(a) Sketch a possible prior density for $\theta$ if you believe that $\theta$ is unlikely to be smaller than 0.1, and that the most likely value of $\theta$ is 0.8.

(b) Sketch a possible prior density for $\theta$ if you believe that the most likely value of $\theta$ is 0.8, but that $\theta$ is unlikely to be smaller than 0.65 or larger than 0.9. Compare your prior density with the prior density you sketched in part (a).

(c) Sketch a possible prior density for $\theta$ if you are almost certain that $\theta$ is 0.8.

(d) Sketch the prior density for $\theta$ if you believe that all values between 0 and 1 are equally likely.

(e) Which of the four prior densities in parts (a) to (d) is the weakest? Order the four densities from weak to strong.

## 3.2 Conditional distributions

In Section 2, you saw that updating beliefs about a proposition involves the use of conditional probabilities. Updating beliefs about a parameter $\theta$ requires the use of conditional distributions. These are discussed briefly in this subsection. The idea of a conditional distribution is introduced in Example 3.3 for a situation where the random variables are discrete.

### Example 3.3 Conditional distribution of CHD, given gender

Data on deaths from coronary heart disease (CHD) by gender were introduced in Example 2.1. The variable $X$ represents the gender of a person who died, and $Y$ indicates whether or not a person died from CHD. In Example 2.1, the conditional probability that $Y = \text{yes}$, given that $X = \text{male}$, was estimated to be 0.2236. Thus, as $Y$ takes only two possible values, yes and no,

$$\begin{aligned} P(Y = \text{no}|X = \text{male}) &= 1 - P(Y = \text{yes}|X = \text{male}) \\ &\simeq 1 - 0.2236 = 0.7764. \end{aligned}$$

The two probabilities $P(Y = \text{yes}|X = \text{male})$ and $P(Y = \text{no}|X = \text{male})$ determine the **conditional distribution** of $Y$, given $X = \text{male}$. ♦

### Activity 3.3 Conditional distribution of gender, given CHD

(a) In Activity 2.2, you estimated the probability $P(X = \text{male}|Y = \text{yes})$ to be 0.5488. Hence obtain the conditional distribution of $X$, given $Y = \text{yes}$.

(b) In Example 2.2, the probability $P(X = \text{male})$ was estimated to be 0.4752. Hence verify that the distribution of $X$ is not the same as the conditional distribution of $X$, given $Y = \text{yes}$.

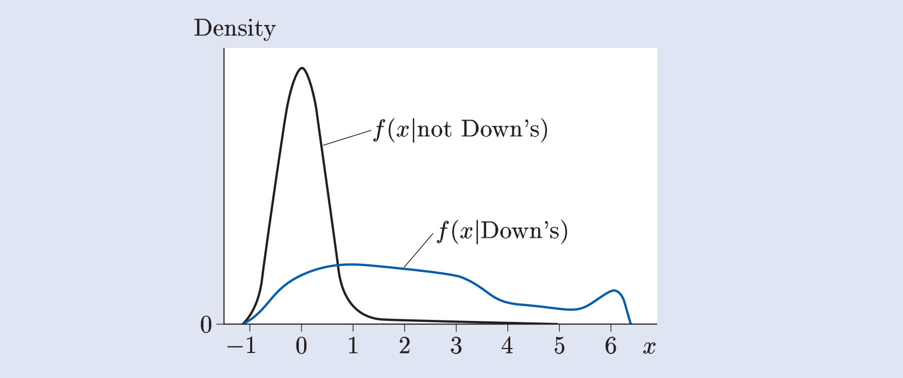
For discrete random variables $X$ and $Y$, the conditional distribution of $Y$, given $X = x_0$, is a discrete distribution, so it has an associated probability mass function; this is denoted $p(y|X = x_0)$. Conditional distributions are defined analogously for continuous random variables $X$ and $Y$. In this case, the conditional distribution of $Y$, given $X = x_0$, has a probability density function which is denoted $f(y|X = x_0)$.

Activity 3.4 involves the conditional distribution of a continuous random variable, given the value of a discrete random variable.

### Activity 3.4 Nuchal translucency measurements, given Down's

When a foetus is aged between 11 and 14 weeks, a nuchal translucency scan can be carried out; this measures the fluid at the back of the baby's neck (the nuchal translucency). This measurement is associated with the length from the crown to the rump of the foetus. The crown–rump measurement allows doctors to estimate the nuchal translucency thickness. A single measurement of a variable $X$ is obtained, where $X$ is the nuchal translucency thickness observed from the scan minus the expected nuchal translucency thickness estimated from the baby's crown–rump measurement.

The distribution of $X$ for foetuses with Down's syndrome is quite different from the distribution for foetuses without Down's syndrome. Hence there are two conditional distributions of interest: the conditional distribution of $X$, given that the foetus has Down's syndrome, and the conditional distribution of $X$, given that the foetus does not have Down's syndrome. Their p.d.f.s may be written $f(x|\text{Down's})$ and $f(x|\text{not Down's})$. Plots of these two conditional probability density functions are given in Figure 3.3.



Source: Kypros H. Nicolaides (2004) 'The 11–13$^{+6}$ weeks scan', Fetal Medicine Foundation, London.

*Figure 3.3* Densities of $X$ for foetuses with and without Down's syndrome

(a) Is a value of $X$ less than 0 more likely in a foetus with Down's syndrome or in a foetus without Down's syndrome? Explain your answer.

(b) Is a value of $X$ greater than 2 more likely in a foetus with Down's syndrome or in a foetus without Down's syndrome? Explain your answer.

## 3.3 The likelihood function

Suppose that the random variable $X$ has some distribution with unknown parameter $\theta$. If it were known that the value of $\theta$ is $\theta_0$, then the distribution of $X$ would be known exactly. If $X$ is discrete then, *conditional on* $\theta = \theta_0$, the (conditional) probability mass function $p(x|\theta = \theta_0)$ can be written down. Similarly, if $X$ is continuous, the conditional probability density function $f(x|\theta = \theta_0)$ can be written down.

In the Bayesian framework, $\theta$ has a distribution, so statements such as 'conditional on $\theta = \theta_0$' make sense.

This is illustrated in Example 3.4 for a situation where $X$ is discrete.

### Example 3.4 Are politicians truthful?

In Example 3.2, $\theta$ was defined to be the proportion of individuals in Great Britain who trust politicians to tell the truth. In the opinion poll mentioned in that example, 2004 adults were interviewed. However, to simplify the calculations, consider just the first 50 adults interviewed. The number of individuals who said that they trusted politicians to tell the truth is a single observation on a random variable $X$ which can be modelled by the binomial distribution $B(50, \theta)$. Hence, if it is known that $\theta$ is some value $\theta_0$, then the distribution of $X$ is also known. Its p.m.f. is

$$p(x|\theta = \theta_0) = \binom{50}{x} \theta_0^x (1 - \theta_0)^{50-x}, \quad x = 0, 1, \ldots, 50.$$

For example, for $\theta_0 = 0.3$,

$$P(X = x|\theta = 0.3) = \binom{50}{x} 0.3^x 0.7^{50-x}. \quad ♦$$

### Activity 3.5 Are politicians truthful?

Write down the probability mass function of $X$, given that $\theta = 0.2$, where $X$ and $\theta$ are as defined in Example 3.4.

Of course, in Example 3.4 and Activity 3.5, it is not known that $\theta$ is 0.3 or 0.2, or indeed what its value is. In general, the value of a parameter $\theta$ of interest is not known. However, if $X$ is discrete, the conditional p.m.f. $p(x|\theta = \theta_0)$ for different values of $\theta_0$ can be used to help decide which values of $\theta$ are most likely for the observed data, and which are unlikely. Similarly, if $X$ is continuous, the conditional p.d.f. $f(x|\theta = \theta_0)$ can be used.

First, consider the case where $X$ is modelled by a discrete distribution with unknown parameter $\theta$. Suppose that the single data value $X = x_1$ has been observed. For two possible values of $\theta$, $\theta_1$ and $\theta_2$, the associated conditional probabilities are $P(X = x_1|\theta = \theta_1)$ and $P(X = x_1|\theta = \theta_2)$. Suppose that

$$P(X = x_1|\theta = \theta_1) > P(X = x_1|\theta = \theta_2).$$

This means that the probability of observing $x_1$ is larger if $\theta = \theta_1$ than if $\theta = \theta_2$. Therefore, the value $\theta = \theta_1$ explains the data better than the value $\theta = \theta_2$. In other words, as $x_1$ has been observed, $\theta_1$ is a 'more likely' value for $\theta$ than $\theta_2$.

Similarly, if $X$ is modelled by a continuous distribution with unknown parameter $\theta$ then, given the observed value $x_1$ of $X$,

$$f(x_1|\theta = \theta_1) > f(x_1|\theta = \theta_2)$$

implies that $\theta_1$ is a more likely value for $\theta$ than $\theta_2$.

### Example 3.5 Which value of $\theta$ is more likely?

In the poll described in Example 3.4, 22% of the 2004 adults interviewed said that they trusted politicians to tell the truth. Suppose that for the first 50 individuals interviewed, the observation $X = 11$ is obtained.

Consider two possible values of the parameter $\theta$, $\theta = 0.2$ and $\theta = 0.3$. For the observed value $X = 11$,

$$P(X = 11|\theta = 0.2) = \binom{50}{11} 0.2^{11} 0.8^{39} \simeq 0.127,$$

$$P(X = 11|\theta = 0.3) = \binom{50}{11} 0.3^{11} 0.7^{39} \simeq 0.060.$$

Since

$$P(X = 11|\theta = 0.2) > P(X = 11|\theta = 0.3),$$

$\theta$ is more likely to be 0.2 than 0.3, for the observation $X = 11$. ♦

### Activity 3.6 Which value of $\theta$ is more likely?

Suppose that in Example 3.5, instead of $X = 11$, the observation $X = 14$ was made. In this case, $P(X = 14|\theta = 0.2) \simeq 0.050$ and $P(X = 14|\theta = 0.3) \simeq 0.119$. Which value of $\theta$ is more likely, given the observation $X = 14$, $\theta = 0.2$ or $\theta = 0.3$? Explain your answer.

Given an observation $x$ on a discrete random variable $X$, the value of the conditional p.m.f. $p(x|\theta = \theta_0)$ can be calculated for each possible value $\theta_0$ of $\theta$. Since a value is defined for each possible value of $\theta$, these values can be viewed as values of a function of $\theta$, which can be written $p(x|\theta)$. This function is called the **likelihood function**, or simply the **likelihood**. It represents how likely the possible values of $\theta$ are for the observed data $x$.

When viewed as a function of $x$, the notation $p(x|\theta)$ represents the conditional p.m.f. of $X$ given $\theta$; when viewed as a function of $\theta$, $p(x|\theta)$ is the likelihood function given the observation $X = x$. Therefore, to emphasize that the likelihood is a function of $\theta$, and to distinguish it from the probability mass function $p(x|\theta)$, which is a function of $x$, the likelihood for $\theta$ is denoted $L(\theta)$.

### *Example 3.6 The likelihood function*

In Example 3.5, the observed value of $X$ was 11. For this observed value, the likelihood function is

$$L(\theta) = P(X = 11|\theta) = \binom{50}{11}\theta^{11}(1-\theta)^{39}.$$

This function of $\theta$ is defined for all values of $\theta$ in the interval $[0, 1]$. A plot of the likelihood function is shown in Figure 3.4. Notice that although $X$ is a discrete random variable with discrete probability mass function $p(x|\theta)$, the likelihood function is continuous, because it is a function of $\theta$, and $\theta$ can take any value in the interval $[0, 1]$. The likelihood function has a peak when $\theta = 0.22$, the observed proportion of adults who said they trust politicians to tell the truth. This means that when 22% of the replies are observed to be 'yes', the most likely value of $\theta$ is 0.22. For values of $\theta$ greater than 0.5, the likelihood is close to 0. Therefore it is not likely that the true value of $\theta$ is greater than 0.5 when only a small proportion (22%) of 'yes' replies has been observed. ♦

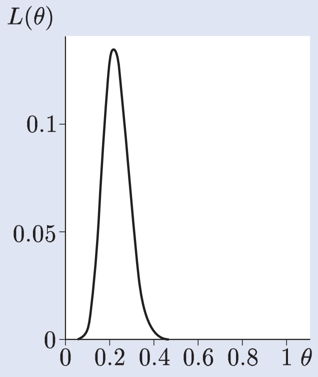


*Figure 3.4* The likelihood function for $\theta$ given $X = 11$

### *Activity 3.7 Calculating a simple likelihood function*

In Example 3.1, $\theta$ was the probability of developing cancer at Slater School. For simplicity, it was assumed that there are only four possible values of $\theta$: 0.03, 0.04, 0.05 and 0.06.

In reality, the set of possible values of $\theta$ is almost always a continuous interval, so the likelihood function will be continuous.

There were 145 staff at Slater School. Let the variable $X$ represent the number of staff who developed cancer at the school.

(a) What are the four possible probability models for $X$? Write down the p.m.f. of the model corresponding to $\theta = 0.03$.

(b) There were 8 cases of cancer among the staff at Slater School. The values of the likelihood function for the four possible values of $\theta$ given the observed value $X = 8$ are shown in Table 3.2.

*Table 3.2* Values of the likelihood function

| $\theta$ | 0.03 | 0.04 | 0.05 | 0.06 |
|---|---|---|---|---|
| $L(\theta)$ | 0.040 | 0.097 | 0.138 | 0.139 |

Of the four values of $\theta$, which is the most likely given the observation $X = 8$? Which is the least likely value for $\theta$?

The likelihood function $L(\theta)$ is defined analogously when $X$ is a continuous random variable: the only difference is that the conditional p.m.f. $p(x|\theta = \theta_0)$ is replaced by the conditional p.d.f. $f(x|\theta = \theta_0)$ in its derivation. Given an observation $X = x$, the likelihood is the function of $\theta$ defined by $L(\theta) = f(x|\theta)$.

More generally, in a statistical inference problem, the data consist of $n$ independent observations $x_1, \dots, x_n$ on $X$. In this case, the likelihood is of the following form:

$$L(\theta) = p(\text{data}|\theta) = p(x_1|\theta) \times \cdots \times p(x_n|\theta) \quad \text{if } X \text{ is discrete,}$$
$$L(\theta) = f(\text{data}|\theta) = f(x_1|\theta) \times \cdots \times f(x_n|\theta) \quad \text{if } X \text{ is continuous.}$$

Whatever form the data take, the general form of the likelihood function $L(\theta)$ is the same, namely $p(\text{data}|\theta)$ or $f(\text{data}|\theta)$, viewed as a function of $\theta$, over all possible values of $\theta$. The details of calculating likelihood functions will be omitted: you will not be expected to calculate likelihood functions in this course. However, the likelihood function is a key element of a Bayesian analysis.

The main ideas concerning the likelihood function that are used in this book are summarized in the following box.

**The likelihood function**

The likelihood function for an unknown parameter $\theta$ is given by

$$L(\theta) = \begin{cases} p(\text{data}|\theta) & \text{for discrete data,} \\ f(\text{data}|\theta) & \text{for continuous data,} \end{cases} \tag{3.1}$$

where $p(\text{data}|\theta)$, or $f(\text{data}|\theta)$, is viewed as a function of $\theta$, over all possible values for $\theta$.

For two possible values of $\theta$, $\theta_1$ and $\theta_2$, if

$$L(\theta_1) > L(\theta_2),$$

then for the observed data, $\theta_1$ is a more likely value of $\theta$ than $\theta_2$.

## 3.4 Posterior distributions

In Section 2, Bayes' theorem was used to update $P(A)$, the prior probability of an event $A$, given the information that event $B$ has occurred. This was done by calculating the conditional probability $P(A|B)$, which is the posterior probability of $A$, given $B$.

Now suppose that we have an unknown parameter $\theta$ and its prior distribution representing our subjective beliefs about the possible values for $\theta$. We want to use some data to make inferences about $\theta$. So after observing data, we wish to update the prior distribution for $\theta$, taking the data into consideration. This requires finding the conditional distribution of $\theta$ given the observed data. As the conditional distribution of $\theta$ given the data represents our updated subjective beliefs about $\theta$ after observing the data, it is called the **posterior distribution**, or simply the **posterior**, for $\theta$. The associated p.d.f. is called the **posterior density**.

Possible values of $\theta$ almost always lie in a continuous interval, so both the prior and posterior distributions for $\theta$ are continuous with associated p.d.f.s $f(\theta)$ and $f(\theta|\text{data})$, respectively. The posterior density can be found using **Bayes' theorem for distributions**, which has the following general form:

$$f(\theta|\text{data}) = \begin{cases} \dfrac{p(\text{data}|\theta)f(\theta)}{p(\text{data})} & \text{for discrete data,} \\ \dfrac{f(\text{data}|\theta)f(\theta)}{f(\text{data})} & \text{for continuous data.} \end{cases} \tag{3.2}$$

The denominator is not a function of $\theta$; it is constant whatever the true value of $\theta$. Since interest lies primarily in the *shape* of the posterior density, the precise value of this constant is not of interest. Therefore (3.2) is often written as

$$f(\theta|\text{data}) \propto \begin{cases} p(\text{data}|\theta)f(\theta) & \text{for discrete data,} \\ f(\text{data}|\theta)f(\theta) & \text{for continuous data.} \end{cases} \tag{3.3}$$

The symbol '$\propto$' means 'is proportional to'.

Since, from (3.1), $f(\text{data}|\theta)$ (or $p(\text{data}|\theta)$) is the likelihood of $\theta$, (3.3) may be written as

$$f(\theta|\text{data}) \propto L(\theta)f(\theta). \tag{3.4}$$

Bayesian inference is based on the properties of the posterior distribution.

Since (3.4) describes (up to some constant) the relationship between the posterior density $f(\theta|\text{data})$, the prior density $f(\theta)$ and the likelihood $L(\theta)$, it is often written in words as

posterior ∝ likelihood × prior.

The posterior distribution combines two sources of information about $\theta$ — the prior distribution, which represents prior subjective beliefs about $\theta$, and the likelihood function, which represents the information about $\theta$ provided by the data. Therefore, in principle, the posterior distribution contains all the available information. It is used for estimating parameters, making predictions, or any other purpose for which the data were gathered. The relationship between the posterior, the prior and the likelihood, which lies at the heart of Bayesian inference, is summarized in the following box.

### Bayesian inference

Bayesian inference is based on the posterior distribution for $\theta$, given the data, denoted $f(\theta|\text{data})$. This is obtained from the prior density $f(\theta)$ and the likelihood $L(\theta)$ using the following expression:

$$f(\theta|\text{data}) \propto L(\theta)f(\theta),$$

or, in words,

posterior ∝ likelihood × prior.

In a Bayesian analysis, the prior distribution, which encapsulates prior beliefs, is combined with the likelihood, which summarizes the information contained in observed data, using Bayes' theorem, to give the posterior distribution. This is used to calculate estimates and make inferences. This process is represented in Figure 3.5.

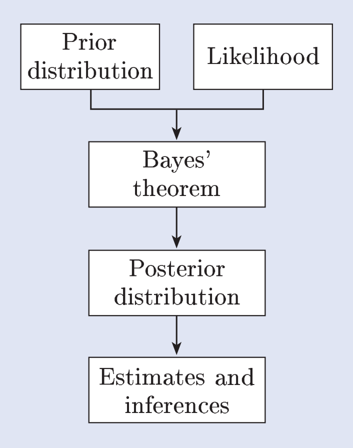


*Figure 3.5* The process of a Bayesian analysis

## *Example 3.7 Bayesian analysis*

A possible prior for $\theta$, the proportion of individuals in Great Britain who trust politicians to tell the truth, was shown in Figure 3.1. The likelihood function for $\theta$ after observing 11 'yes' replies from 50 interviewees ($X = 11$) was given in Figure 3.4 (see Example 3.6).

See Example 3.2.

Figure 3.6 shows the prior, the likelihood and the resulting posterior for $\theta$.

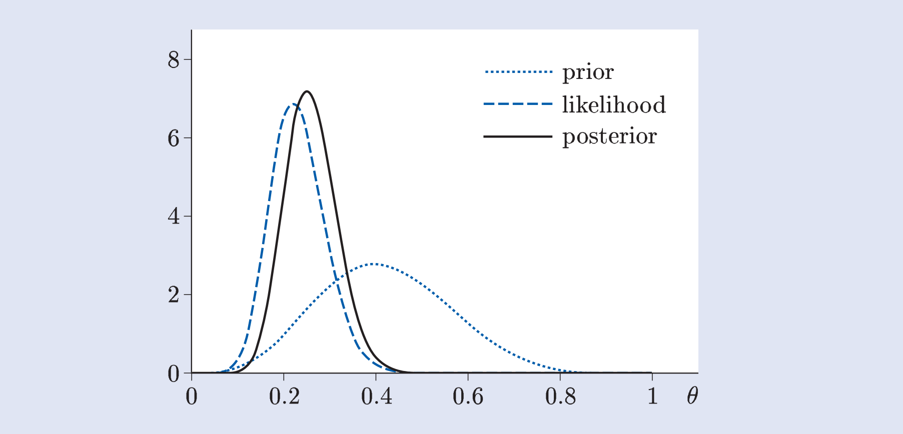


*Figure 3.6* Prior, likelihood and posterior for $\theta$

Since the prior and the posterior are distributions, the area under their densities is 1. In Figure 3.6, the likelihood has been scaled so that the area underneath it is also 1. This makes it easier to see how the prior and the likelihood combine to produce the posterior. Notice that the posterior lies between the prior and the likelihood: since the posterior combines information from the prior and from the data, as represented by the likelihood, it will always lie between the prior and the likelihood, as in Figure 3.6. Notice also that the posterior is much closer to the likelihood than to the prior, because the amount of information contained in the data is much greater than that contained in the prior. In this sense, the likelihood is stronger than the prior.

From now on, all plots of likelihoods will be scaled in this way.

The posterior density in Figure 3.6 summarizes all information about $\theta$ after observing $X = 11$. Prior to observing the data, it was believed that the most likely value of $\theta$ was 0.4 — the peak of the prior is at 0.4 — and that the value of $\theta$ was between 0.05 and 0.9. The posterior density is much narrower than the prior, so after observing the data, we are more confident about the value of $\theta$. The posterior density represents the belief that $\theta$ lies between 0.1 and about 0.45. The most likely value of $\theta$ is now believed to be 0.25. ♦

Since the posterior density $f(\theta|\text{data})$ represents what is known about $\theta$ after the data have been observed, it is used to make inferences about $\theta$. In any Bayesian analysis, it is always a good idea to obtain a plot of $f(\theta|\text{data})$, as this gives an overall graphical summary of the posterior information about $\theta$. Measures of location and spread of the posterior density can also be very useful.

The process of obtaining the posterior distribution and using it for inference is generally referred to as **prior to posterior analysis**. In Section **??**, you will carry out a prior to posterior analysis for a proportion $\theta$.

## Summary of Section 3

In this section, a framework for Bayesian inference has been introduced. The prior distribution for a parameter $\theta$, which has p.d.f. $f(\theta)$, summarizes beliefs about the value of a parameter $\theta$ before data are observed. The likelihood function for $\theta$, which is denoted $L(\theta)$, summarizes the information about $\theta$ contained in the observed data. The posterior distribution, which has p.d.f. $f(\theta|\text{data})$, represents what is known about $\theta$ after the data have been observed. The posterior density combines the two information sources about $\theta$, the prior and the likelihood, via the expression

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

# Exercise on Section 3

## Exercise 3.1 Explaining the posterior

The prior, likelihood and posterior for three different prior to posterior analyses for a proportion $\theta$ are shown in Figure 3.7.

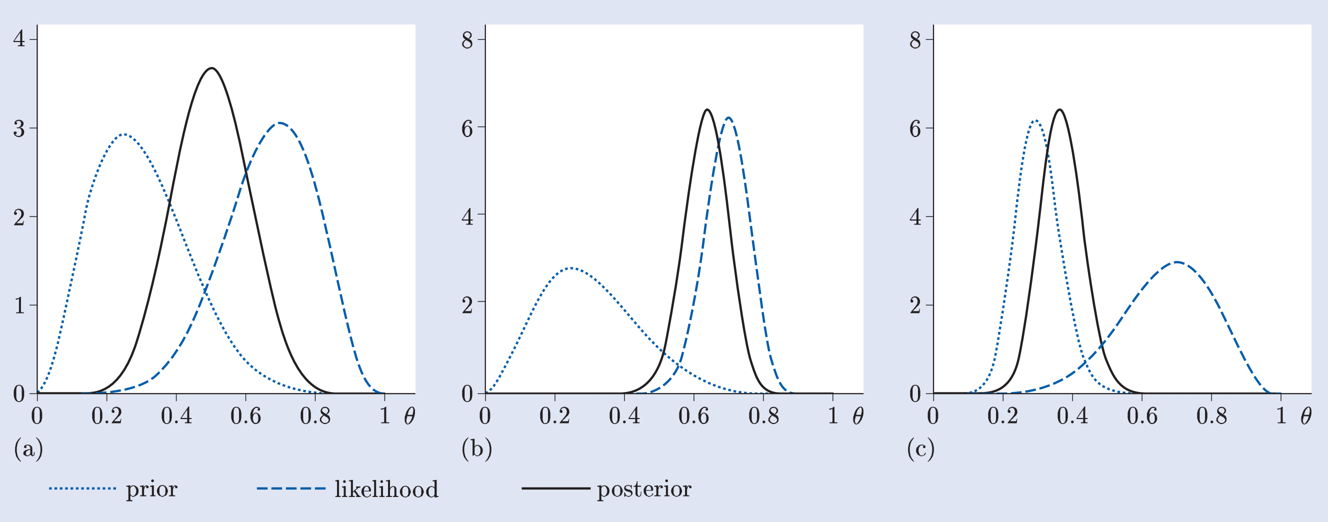


*Figure 3.7* The prior, likelihood and posterior for a proportion $\theta$: three analyses

For each prior to posterior analysis, explain why the posterior looks like it does in relation to the prior and the likelihood.

# Solutions to Activities

## Solution 1.1

$$\begin{aligned} &P(\text{even number}) \\ &= \frac{\text{number of faces on the die with even numbers}}{\text{number of faces on the die}} \\ &= \frac{3}{6} = \frac{1}{2}. \end{aligned}$$

## Solution 1.2

**(a)** Relative frequencies calculated from the data can be used to estimate these probabilities.

**(i)** The probability that a man dies of CHD can be estimated as

$$\begin{aligned} \frac{\text{number of male deaths from CHD}}{\text{number of male deaths}} &= \frac{64\,473}{288\,332} \\ &\simeq 0.2236. \end{aligned}$$

**(ii)** The probability that a woman dies of CHD can be estimated as

$$\begin{aligned} \frac{\text{number of female deaths from CHD}}{\text{number of female deaths}} &= \frac{53\,003}{318\,463} \\ &\simeq 0.1664. \end{aligned}$$

**(b)** These estimates are based on large amounts of data from 2002. Assuming that the probability of death from CHD has not changed since 2002, these estimates will be very good. Of course, it is possible that the probability of death from CHD may change over time, so it is important to bear in mind that the further in the future the year is, the less reliable these estimates may be.

## Solution 1.3

It is easier to recall words which begin with r (*r*un, *r*abbits, ...) than words with r as the third letter (pa*r*k, bi*r*d, ...). Consequently, most people judge it more likely that an English word begins with r than that r is the third letter, that is, $p_1 > p_3$. In fact, the reverse is true: $p_3 > p_1$. This is an example of possible probability estimation bias, where the person making the estimate links the probability to the frequency with which they can recall occurrences, rather than the frequency with which the event actually occurs.

(Adapted from: Tversky, A. and Kahneman, D. (1973) Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, **5**, 207–232.)

## Solution 2.1

**(a)** **(i)** From Table 2.2,

$$P(\text{Down's}|\text{mother's age} = 25\text{ years}) = 1/1352$$

**(ii)** From Table 2.2,

$$P(\text{Down's}|\text{mother's age} = 40\text{ years}) = 1/97.$$

**(b)** To compare the numbers of pregnancies affected by Down's syndrome, the total numbers of pregnancies in women aged 25 and in women aged 40 are required. This information is not available, so it is not possible to conclude that more foetuses with Down's syndrome occur in pregnant women aged 40 than in pregnant women aged 25.

## Solution 2.2

**(a)** Using the data in Table 2.1 gives the estimate

$$\begin{aligned} &P(X = \text{male}|Y = \text{yes}) \\ &= \frac{\text{number of male deaths from CHD}}{\text{total number of deaths from CHD}} \\ &= \frac{64\,473}{117\,476} \simeq 0.5488. \end{aligned}$$

**(b)** Using Formula (2.1) gives

$$\begin{aligned} &P(X = \text{male}|Y = \text{yes}) \\ &= \frac{P(Y = \text{yes and } X = \text{male})}{P(Y = \text{yes})} \\ &\simeq \frac{0.1063}{0.1936} \simeq 0.5491. \end{aligned}$$

This is the same to three decimal places as the estimate produced using the data in Table 2.1 directly. (The slight discrepancy is due to rounding error.)

## Solution 2.3

**(a)** From Table 2.1, an estimate of the conditional probability is given by

$$P(Y = \text{yes}|X = \text{female}) = \frac{53\,003}{318\,463} \simeq 0.1664.$$

**(b)** Using Formula (2.4),

$$\begin{aligned} &P(Y = \text{yes}) \\ &= P(Y = \text{yes}|X = \text{male}) \times P(X = \text{male}) \\ &\quad + P(Y = \text{yes}|X = \text{female}) \times P(X = \text{female}). \end{aligned}$$

Since $P(X = \text{male}) + P(X = \text{female}) = 1$,

$$\begin{aligned} P(X = \text{female}) &= 1 - P(X = \text{male}) \\ &\simeq 1 - 0.4752 \\ &= 0.5248. \end{aligned}$$

Hence an estimate of $P(Y = \text{yes})$ is given by

$$\begin{aligned} P(Y = \text{yes}) &\simeq 0.2236 \times 0.4752 + 0.1664 \times 0.5248 \\ &\simeq 0.1936. \end{aligned}$$

This is the same as the estimate found directly from the contingency table in Example 2.1.

## Solution 2.4

In Example 2.6, you saw that

$$P(\text{Adams guilty}|M) \simeq \frac{1 \times \frac{1}{200\,000}}{P(M)},$$

where $P(M)$ is given by

$$P(M|\text{Adams guilty}) \times P(\text{Adams guilty}) \\ + P(M|\text{Adams not guilty}) \times P(\text{Adams not guilty}).$$

Using the prosecution's estimate of $\frac{1}{200\,000\,000}$ for $P(M|\text{Adams not guilty})$ gives

$$\begin{aligned} P(M) &\simeq 1 \times \frac{1}{200\,000} + \frac{1}{200\,000\,000} \times \left(1 - \frac{1}{200\,000}\right) \\ &= \frac{200\,000\,000}{200\,000\,000 \times 200\,000} + \frac{199\,999}{200\,000\,000 \times 200\,000} \\ &= \frac{200\,199\,999}{200\,000\,000 \times 200\,000}. \end{aligned}$$

Thus

$$\begin{aligned} &P(\text{Adams guilty}|M) \\ &\simeq \left(1 \times \frac{1}{200\,000}\right) \Big/ \frac{200\,199\,999}{200\,000\,000 \times 200\,000} \\ &= \frac{200\,000\,000}{200\,199\,999} \\ &\simeq 0.999. \end{aligned}$$

Since the probability is very close to 1, this could well be evidence of guilt 'beyond all reasonable doubt'. Thus, based on the prosecution's estimate of $P(M|\text{Adams not guilty})$ and the prosecution's evidence $M$, the judge and jury may well believe Adams to be guilty. However, evidence from the defence has not yet been considered.

## Solution 2.5

In the solution to Activity 2.4, you calculated $P(\text{Adams guilty}|M)$ to be 0.999 using the prosecution's estimate of $P(M|\text{Adams not guilty})$. This probability can be updated after observing $E_1$ and $E_2$, using the method of Example 2.7, but with $P(\text{Adams guilty}|M) = 0.999$.

First, updating for $E_1$, (2.10) gives

$$\begin{aligned} P(E_1|M) &\simeq 0.1 \times 0.999 + 0.9 \times (1 - 0.999) \\ &= 0.1008. \end{aligned}$$

Then (2.9) leads to

$$P(\text{Adams guilty}|E_1, M) = \frac{0.1 \times 0.999}{0.1008} \simeq 0.991.$$

Secondly, updating for $E_2$, (2.12) becomes

$$\begin{aligned} P(E_2|E_1, M) &\simeq 0.25 \times 0.991 + 0.5 \times (1 - 0.991) \\ &= 0.252\,25. \end{aligned}$$

Using this value in (2.11) leads to

$$P(\text{Adams guilty}|E_2, E_1, M) \simeq \frac{0.25 \times 0.991}{0.252\,25} \simeq 0.98.$$

With the prosecution's estimate of $P(M|\text{Adams not guilty}) = 1/200\,000\,000$, the posterior probability of guilt after all the evidence $(M, E_1, E_2)$ has been considered is very close to 1. This may well be judged evidence of guilt 'beyond all reasonable doubt'.

## Solution 3.1

The prior beliefs about $\theta$ are expressed through the prior distribution. From the prior distribution, the value $\theta = 0.03$ has the largest probability, so the value 0.03 is the most likely value of $\theta$.

## Solution 3.2

**(a)** A sketch of a possible prior density is shown in Figure S.1.

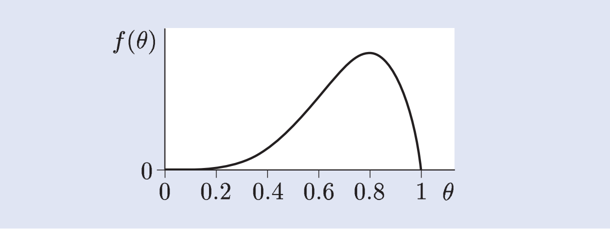


*Figure S.1* A possible prior density for $\theta$

Notice that the density has a peak at $\theta = 0.8$, as 0.8 is believed to be the most likely value for $\theta$. Also, the density lies between the values 0.1 and 1 to reflect the belief that $\theta$ is unlikely to be smaller than 0.1.
Since 0.8, the most likely value for $\theta$, is not equidistant between 0.1 and 1, this density is skewed.

**(b)** A sketch of a possible prior density is shown in Figure S.2.

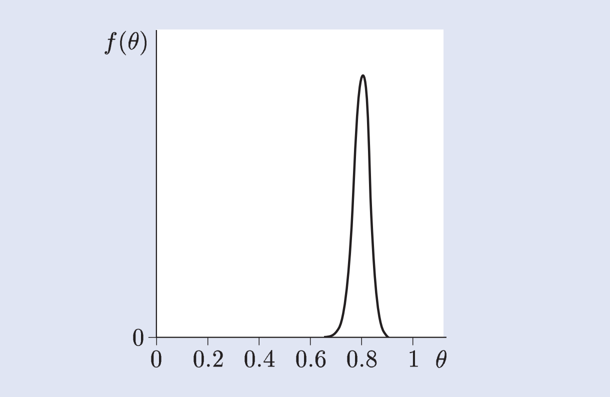


*Figure S.2* A possible prior density for $\theta$

Both this prior density and the one for part (a) have a peak at $\theta = 0.8$, as 0.8 is believed to be the most likely value for $\theta$ in each case. However, the prior density in Figure S.2 is narrower and taller than the prior density in Figure S.1 because the range of likely values for $\theta$ is smaller for the prior density in Figure S.2, but the area under the curve is 1 for both densities. Both densities are skewed.

**(c)** A sketch of a possible prior density is shown in Figure S.3.

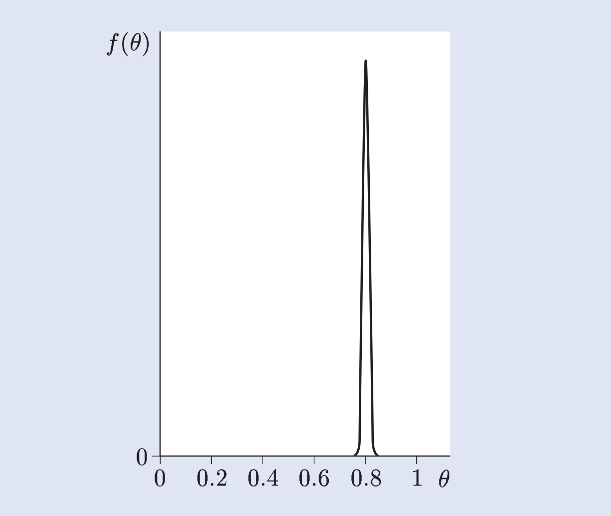


*Figure S.3* A possible prior density for $\theta$

Notice that the entire density is concentrated around the value 0.8 because you are almost certain that $\theta$ is 0.8.

**(d)** The sketch of the prior density is shown in Figure S.4.

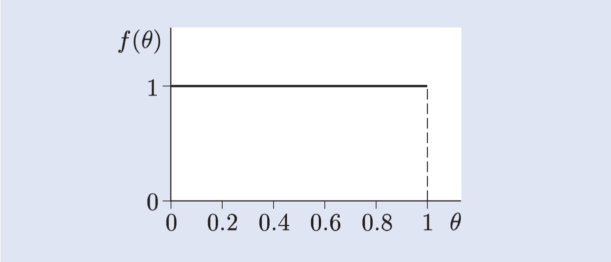


*Figure S.4* The prior density for $\theta$

Notice that all possible values of $\theta$ between 0 and 1 are equally likely: the density is flat. You might use a prior density like this if you have no idea what the value of $\theta$ is.

**(e)** The prior density in Figure S.4 is the weakest as it represents the fact that no information is available about $\theta$. The prior density in Figure S.1 is also fairly weak, as the range of likely values includes almost all values between 0 and 1. The prior density in Figure S.2 is stronger, representing greater confidence that $\theta$ lies in a narrower range of values. The prior density in Figure S.3 is the strongest prior density, as this represents virtual certainty that 0.8 is the value of $\theta$. In reality, it is unlikely that such a strong prior density would represent your belief about $\theta$.

## Solution 3.3

**(a)** Since $X$ can take only two values, male and female,

$$\begin{aligned}&P(X = \text{female}|Y = \text{yes})\\&= 1 - P(X = \text{male}|Y = \text{yes})\\&\simeq 1 - 0.5488 = 0.4512.\end{aligned}$$

The probabilities $P(X = \text{male}|Y = \text{yes})$ and $P(X = \text{female}|Y = \text{yes})$ give the conditional distribution of $X$, given $Y = \text{yes}$.

**(b)** Since $P(X = \text{male}) \simeq 0.4752$, it follows that $P(X = \text{female}) \simeq 0.5248$. Thus the distribution of $X$ is indeed different from the conditional distribution of $X$, given $Y = \text{yes}$.

## Solution 3.4

**(a)** A value of $X$ less than 0 is more likely in a foetus without Down's syndrome: the value 0 is close to the median of $f(x|\text{not Down's})$, but well below the median of $f(x|\text{Down's})$.

**(b)** A value of $X$ greater than 2 is more likely in a foetus with Down's syndrome. The value of $X$ is greater than 2 in very few foetuses without Down's syndrome, whereas such values are quite common in foetuses with Down's syndrome.

## Solution 3.5

When $\theta = 0.2$, the p.m.f. of $X$ is

$$P(X = x|\theta = 0.2) = \binom{50}{x} 0.2^x 0.8^{50-x}.$$

## Solution 3.6

If $X = 14$, then the value 0.3 is more likely for $\theta$ than 0.2, because

$$P(X = 14|\theta = 0.3) > P(X = 14|\theta = 0.2).$$

## Solution 3.7

**(a)** As the probability of developing cancer at Slater School is assumed to be the same for all 145 staff, and staff are assumed to develop cancer independently, $X$ can be modelled by the binomial distribution $B(145, \theta)$. The parameter $\theta$ can take only four possible values: 0.03, 0.04, 0.05 or 0.06. Thus there are just four possible models: the binomial models $B(145, 0.03)$, $B(145, 0.04)$, $B(145, 0.05)$ and $B(145, 0.06)$.

The p.m.f. corresponding to the $B(145, 0.03)$ model is

$$P(X = x|\theta = 0.03) = \binom{145}{x} 0.03^x 0.97^{145-x}.$$

**(b)** The largest probability of observing $X = 8$ is when $\theta = 0.06$, so this is the most likely value of $\theta$. Similarly, the least likely value of $\theta$ is 0.03.

# Solutions to Exercises

## Solution 1.1

**(a)** The relative frequency can be used to estimate the probability, so an estimate is given by

$$\frac{\text{number of mothers who pack crisps every day}}{\text{number of mothers questioned}}$$
$$= \frac{374}{720} \simeq 0.52.$$

**(b)** Using the relative frequency, an estimate is given by

$$\frac{\text{number of children who bring crisps every day that week}}{\text{number of children in the class}}$$
$$= \frac{19}{28} \simeq 0.68.$$

**(c)** The estimate in part (a) is likely to be more reliable for two reasons. First, it is based on far more data. Secondly, since the estimate in part (b) is based on data from a single class, it may not be representative of the UK as a whole — for example, the children in the class may all be of a similar social background and this could affect the contents of their lunchboxes. (Also, some of the children in part (b) may have packed their own lunchboxes, whereas the implication in part (a) is that the lunchboxes were all packed by the mothers. The data may not be comparable.)

## Solution 2.1

**(a)** **(i)** The test gives a positive result for all people infected with HIV, so

$$P(\text{positive result}|\text{person infected with HIV}) = 1.$$

**(ii)** The test gives a positive result for 0.005% of people not infected with HIV, so

$$P(\text{positive result}|\text{person not infected with HIV})$$
$$= \frac{0.005}{100} = 0.000\,05.$$

**(b)** **(i)** Since 10 in 100 000 women in this low-risk group are infected with HIV, the prior probability that the woman is infected is

$$P(\text{infected}) = \frac{10}{100\,000} = 0.0001.$$

**(ii)** The required probability can be calculated using Bayes' theorem. First, using Formula (2.3),

$$P(\text{infected}|\text{positive})$$
$$= \frac{P(\text{positive}|\text{infected}) \times P(\text{infected})}{P(\text{positive})}$$
$$= \frac{1 \times 0.0001}{P(\text{positive})}.$$

Now applying Formula (2.4) to find $P(\text{positive})$ gives

$$\begin{aligned} &P(\text{positive}) \\ &= P(\text{positive}|\text{infected}) \times P(\text{infected}) \\ &\quad + P(\text{positive}|\text{not infected}) \times P(\text{not infected}) \\ &= 1 \times 0.0001 + 0.000\,05 \times 0.9999 \\ &= 0.000\,149\,995. \end{aligned}$$

Therefore

$$P(\text{infected}|\text{positive}) = \frac{1 \times 0.0001}{0.000\,149\,995} \simeq 0.667.$$

With a posterior probability of 0.667, it is far from certain that a woman in this low-risk group who has a positive test result is infected with HIV.

## Solution 2.2

The posterior probability can be calculated using Bayes' theorem. First, using Formula (2.3),

$$\begin{aligned} &P(\text{in square}|\text{not found}) \\ &= \frac{P(\text{not found}|\text{in square}) \times P(\text{in square})}{P(\text{not found})} \\ &= \frac{(1 - 0.75) \times 0.4}{P(\text{not found})} \\ &= \frac{0.1}{P(\text{not found})}. \end{aligned}$$

Now applying Formula (2.4) to calculate $P(\text{not found})$ gives

$$\begin{aligned} &P(\text{not found}) \\ &= P(\text{not found}|\text{in square}) \times P(\text{in square}) \\ &\quad + P(\text{not found}|\text{not in square}) \times P(\text{not in square}) \\ &= 0.25 \times 0.4 + 1 \times 0.6 \\ &= 0.7. \end{aligned}$$

Hence

$$P(\text{in square}|\text{not found}) = \frac{0.1}{0.7} \simeq 0.143.$$

## Solution 2.3

**(a)** **(i)** If a female carrier has a son, then the probability that the son is affected by haemophilia is 0.5, so

$$\begin{aligned} P(\text{Jack is not affected}|\text{Kate is a carrier}) &= 1 - 0.5 \\ &= 0.5. \end{aligned}$$

**(ii)** If a female is not a carrier, then any son she may have will not be affected, so

$$P(\text{Jack is not affected}|\text{Kate is not a carrier}) = 1.$$

**(b)** Using Bayes' theorem (and abbreviating the names to J and K), Formula (2.3) gives

$$\begin{aligned}&P(\text{Kate is a carrier}|\text{Jack not affected})\\&= \frac{P(\text{J not affected}|\text{K carrier}) \times P(\text{K carrier})}{P(\text{J not affected})}\\&= \frac{0.5 \times 0.5}{P(\text{J not affected})}\\&= \frac{0.25}{P(\text{J not affected})}.\end{aligned}$$

Using Formula (2.4),

$$\begin{aligned}&P(\text{J not affected})\\&= P(\text{J not affected}|\text{K carrier}) \times P(\text{K carrier})\\&\quad + P(\text{J not affected}|\text{K not carrier})\\&\qquad \times P(\text{K not carrier})\\&= 0.5 \times 0.5 + 1 \times 0.5\\&= 0.75.\end{aligned}$$

Hence

$$P(\text{Kate is a carrier}|\text{Jack not affected}) = \frac{0.25}{0.75} = \frac{1}{3}.$$

**(c)** The new posterior probability is

$$P(\text{Kate is a carrier}|\text{Jack and Luke are not affected}).$$

This can be calculated using Bayes' theorem, as given in Formulas (2.7) and (2.8). Abbreviating the names to K, J and L, and using Formula (2.7), gives

$$\begin{aligned}&P(\text{K carrier}|\text{J and L not affected})\\&= P(\text{L not affected}|\text{K carrier, J not affected})\\&\quad \times P(\text{K carrier}|\text{J not affected})\\&\qquad /P(\text{L not affected}|\text{J not affected}).\end{aligned}$$

Since the probability that a carrier has a son who is affected is the same for each son, the probability that Luke is affected does not depend on whether or not Jack is affected. Therefore

$$\begin{aligned}&P(\text{L not affected}|\text{K carrier, J not affected})\\&= P(\text{L not affected}|\text{K carrier}) = 0.5.\end{aligned}$$

From part (b),

$$P(\text{K carrier}|\text{J not affected}) = \tfrac{1}{3}.$$

Hence

$$\begin{aligned}&P(\text{K carrier}|\text{J and L not affected})\\&= \frac{0.5 \times \frac{1}{3}}{P(\text{L not affected}|\text{J not affected})}.\end{aligned}$$

Now using Formula (2.8),

$$\begin{aligned}&P(\text{L not affected}|\text{J not affected})\\&= P(\text{L not affected}|\text{K carrier, J not affected})\\&\qquad \times P(\text{K carrier}|\text{J not affected})\\&\quad + P(\text{L not affected}|\text{K not carrier, J not affected})\\&\qquad \times P(\text{K not carrier}|\text{J not affected}).\end{aligned}$$

Now

$$\begin{aligned}&P(\text{L not affected}|\text{K not carrier, J not affected})\\&= P(\text{L not affected}|\text{K not carrier})\\&= 1,\end{aligned}$$

$$\begin{aligned}&P(\text{K not carrier}|\text{J not affected})\\&= 1 - P(\text{K carrier}|\text{J not affected})\\&= \tfrac{2}{3}.\end{aligned}$$

Hence

$$\begin{aligned}&P(\text{L not affected}|\text{J not affected})\\&= 0.5 \times \tfrac{1}{3} + 1 \times \tfrac{2}{3}\\&= \tfrac{5}{6}.\end{aligned}$$

Therefore the new posterior probability is

$$\frac{0.5 \times \frac{1}{3}}{\frac{5}{6}} = 0.2.$$

## Solution 3.1

The posterior combines the information about $\theta$ contained in the prior with the information about $\theta$ contained in the likelihood.

In Figure 3.7(a), the prior and the likelihood are a similar height and width, so the prior and the likelihood are fairly similar in the strength of information they represent. Thus the posterior lies directly between the prior and the likelihood.

In Figure 3.7(b), the likelihood is strong in comparison to the prior. Therefore the posterior is closer to the likelihood than to the prior.

In Figure 3.7(c), the prior is strong in comparison to the likelihood. Consequently, the posterior is closer to the prior than to the likelihood.

# Index